# EXPERIMENTAL PHYLOGENETICS: A BENCHMARK FOR ANCESTRAL SEQUENCE RECONSTRUCTION

A Thesis
Presented to
The Academic Faculty

by

Ryan Nicole Randall

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Biology

Georgia Institute of Technology
August 2012

# EXPERIMENTAL PHYLOGENETICS: A BENCHMARK FOR ANCESTRAL SEQUENCE RECONSTRUCTION

Approved by:

Dr. Eric Gaucher, Advisor
School of Biology
*Georgia Institute of Technology*

Dr. Soojin Yi
School of Biology
*Georgia Institute of Technology*

Dr. Brian Hammer
School of Biology
*Georgia Institute of Technology*

Date Approved: June 11th, 2012

# ACKNOWLEDGEMENTS

First and foremost I would like to recognize Prof. Eric A. Gaucher and thank him for training me to become a successful scientist and innovative thinker.

A special thanks goes out to all Gaucher lab members. I thank Ziming Zhao for her computational wisdom and for providing a large portion of statistical analysis needed towards the end of this research project. I thank Brian Kwan for his assistance in collecting spectral data and providing me with some PYMOL images. I give gratitude to James Kratzer. He has been with me in the Gaucher lab from start to finish, and he is always ever so helpful. I also thank Josh Stern, Dr. Megan Cole, and Dr. Betul Kacar for their continued support throughout my graduate experience.

I am grateful for my Professors, Dr. Soojin Yi and Dr. Brian Hammer for their helpful guidance as members of my thesis committee. I thank them for their support, encouragement, feedback, and suggestions that helped me in my research.

I owe gratitude to Andrew Shaw and Steve Woodard for providing readily available IBB core equipment for my fluorescent protein analysis. I thank all my family and friends who have continually stood by my side throughout my research experience.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ASR | Ancestral Sequence Reconstruction |
| MP | Maximum Parsimony |
| ML | Maximum Likelihood |
| MSA | Multiple Sequence Alignment |
| mRFP | monomeric Red Fluorescent Protein |
| FP | Fluorescent Protein |
| GFP | Green Fluorescent Protein |
| DNA | Polymerase Chain Reaction |
| TT | True Topology |
| MB | MrBayes |
| AA | Amino Acid |

# SUMMARY

The field of molecular evolution has benefited greatly from the use of ancestral sequence reconstruction as a methodology to better understand the molecular mechanisms associated with functional divergence. The method of ancestral sequence reconstruction has never been experimentally validated despite the method being exploited to generate high profile publications and gaining wider use in many laboratories. The failure to validate such a method is a consequence of 1) our inability to travel back in time to document evolutionary transitions and 2) the slow pace of natural evolutionary processes that prevent biologists from 'witnessing' evolution in action (pace viruses). In this thesis research, we have generated an experimentally known phylogeny of fluorescent proteins in order to benchmark ancestral sequence reconstruction methods. The tips/leaves of the fluorescent protein experimental phylogeny are used to determine the performances of various ASR methods. This is the first example of combining experimental phylogenetics and ancestral sequence reconstruction.

# CHAPTER 1

# PHYLOGENETICS

## Introduction

One of the principle aims of modern evolutionary biologists is to understand patterns of decent, and to use knowledge about these patterns of decent to understand the evolutionary events that have transpired throughout the history of life on Earth. This field is termed systematics/phylogenetics. Scientists in this field study evolutionary relationships using a phylogeny - a branching diagram that shows the evolutionary history among a collection of organisms and/or gene sequences over time. Specifically, in a phylogeny or phylogenetic tree, modern organisms are placed at the leaves of the tree, ancestral organisms occupy the internal nodes, and the branching patterns of the tree denote the evolutionary relationships among the modern and ancestral organisms or genes.

## Importance

Phylogenetic analyses are used in essentially all branches of biology; the applications range from studies on the origin of human populations [1] to predicting influenza's next mode of drug-resistance [2]. Due to the importance of phylogenetic trees in biological inquiry, phylogenetic inference is a vastly growing field, and many phylogenetic algorithms have been developed for inferring phylogenic trees using molecular sequence data, such as DNA and amino acid sequence information [3]. DNA and amino acid sequences can be considered simply as strings composed of either a 4-letter alphabet (A, C, T, and G for DNA) and a 20-letter alphabet (one letter for each amino acid) for a protein sequence

[4]. Phylogenetic reconstruction methods are numerical algorithms. Modern sequences serve as the algorithm's input and, in turn, the algorithm generates a phylogenetic tree showing the evolutionary history based on the data set. The resulting phylogeny serves as a hypothesis given that phylogenetic reconstructions explicitly assume the mechanisms by which the sequences have evolved. Though reconstruction methods make assumptions about evolution, they differ in the details of these assumptions. The methods of phylogenetic inference used in molecular phylogenetics are historically classified into two major groups: maximum likelihood [5] and parsimony methods [6].

Maximum parsimony (MP) assumes the simplest, most parsimonious mode of evolutionary change when given a set of aligned sequences (following Occam's razor). MP will construct a tree and order the tree's internal nodes in such a way as to minimize the total number of DNA/amino acid changes on the tree that would give rise to the modern sequences. MP assumes that the shortest tree is optimal since tree length is positively correlated with the amount of evolutionary change. Thus, the MP tree will always have the minimum number of mutations required to explain the observed site patterns.

Maximum Likelihood (ML) is the probability of the data (alignment), given a tree (with topology and branch lengths specified) and a probabilistic model of evolution: L = P(data | tree and model). Branch lengths represent the expected number of character-state changes along a branch. When a branch is short, there is a relatively low probability of a single change at a site in the sequence occurring along that branch, and an almost negligible probability of more than one change at a site. It is far more likely to have a change occur along a long branch over a short branch. The maximum likelihood approach formulates a probabilistic model of evolution through explicit use of extant sequence data, thus

the tree that maximizes the likelihood of the observed data is the optimal tree. For example, we may know the nucleotide substitution rate among a group of DNA sequences and ML will assume this substitution rate to find the tree that most likely generated the data.

Models of molecular evolution are based on substitution rate matrices. Models vary in the numbers and kinds of parameters used to determine elements in the rate matrix. An explicit model of nucleotide substitution or amino acid replacement is a primary requirement in ML methods [7, 8]. Nucleotide substitution models vary from the double-parameter Kimura 1980 model to the general time reversible model (GTR) [7-9]. These models assume reversible matrices; in other words, they assume that the probability of the forward change over time (e.g., A to G) is equal to the probability of the reverse event (G to A). Other models have been proposed that are based on amino acid sequences, such as the Dayhoff and Jones Taylor Thornton (JTT) models [10], as well as codon sequences [11]. Various approaches exist for choosing models of molecular evolution to incorporate into a phylogenetic analysis, the most common of which is jModelTest for DNA sequences and ProtTest for amino acid sequences [12, 13]. These methods ensure that the best-fit models are neither under parameterized nor over parameterized. The models discussed above have been used to analyze our ancestral reconstructions in section 6.

## Assessment Before 1992

The only means by which we are able to represent the evolutionary connections of organisms and infer the selective forces shaping their evolution is through the use of phylogenetics. Since we so heavily rely on phylogenetic reconstruction to develop such understandings, it is essential that tree

construction methods are in fact valid and accurate. Several methods have been developed and address questions such as whether we know we are building the correct trees and whether we know when MP or ML generates incorrect phylogenies and thus misleads us.

Assessment of phylogenetic reconstruction methods has relied on computational simulations whereby a random sequence is generated and evolved under selected evolutionary models. Numerical simulations assume a particular model of evolution and then generate characters (typically, nucleotide sequences) according to the model and a given phylogeny. Thus, an investigator can generate many replicate data sets under specified conditions in order to compare the performance of competing methods.

Computer simulations attempted to benchmark phylogenetic methods for accurately inferring the correct phylogenetic tree topology and branch length and have shown the pros and cons of various reconstruction methods [14]. Computer simulations have shown the weaknesses of parsimony. For example, computer simulations have determined that parsimony tends to cluster long branches together on the phylogeny, thus leading to incorrect branching patterns when using particular types of sequence data. This phenomenon is known as long branch attraction [15]. Parsimony has also been shown to be inconsistent as more data are accumulated because inferences tend to converge on an incorrect parsimonious solution.

Also, during a simulation, a particular substitution model is defined but there is no way of knowing which substitution model accurately reflects DNA substitutions in real organisms. Since phylogenetic reconstruction accuracy is highly dependent on the data set used, it would be important to simulate sequences in a biological consistent manner – but this is difficult to achieve.

Although simulations provide considerable insight into the effectiveness of various phylogenetic algorithms, they are limited by an incomplete knowledge of biology: all models incorporate untested assumptions about evolutionary processes. Instead of simulating a phylogeny to represent a biological reality, why not create an actual phylogeny of evolved biological organisms? Thus, the field of experimental phylogenetics was born [14].

## Experimental Phylogenetics

An experimental phylogeny is a phylogeny that is built in the laboratory where all of the ancestral characters and evolutionary relationships are known. The primary goal in the field of experimental phylogenetics is to generate branching histories of biological entities in the laboratory for use in testing methods of phylogenetic reconstruction.

Analysis of known phylogenies added a reality check to simulation studies. In 1993, Hillis and Bull evolved virus T7 bacteriophage through serial propagation in the laboratory and generated a known phylogeny providing for the first time experimental support for phylogeny inference methods [14]. Analyzing a known experimental phylogeny from the laboratory, where the phylogeny is evolved

under real biological constraints, rather than modeled conditions, added rigor in evaluating phylogenetic inference performance that computer simulations were unable to provide. The authors' results added legitimacy to methods of phylogenetic estimation.

Since Hillis and Bull's T7 phylogeny, experimental phylogenetics has proved to be a convincing means of understanding basic evolutionary processes. Sousa *et al.* propagated Bacteriophage T7 to test the effects of asymmetry and short branch lengths on phylogenetic inferences [16]. These authors compared a simulated phylogeny to their Bacteriophage T7 phylogney and found that simulations could accommodate many but not all of the problems encountered by phylogenetic *inference* methods. The authors concluded that short internal branches might cause more error (leading to incorrect branching patterns) than previously thought. In another study, Sanson *et al.* generated an experimental phylogeny of *Trypanosoma cruzi* in order to understand the effects neutral substitutions have on phylogenetic inferences. Their results established biochemical experimental support for phylogenies, divergence date estimates, and an irreversible substitution model based on neutrally evolving DNA sequences [17].

Experimental phylogenetics has not been widely adopted by researchers since its birth nearly 20 years ago. Critics of experimental phylogenetics do exist, yet they do not deny the importance of experimental phylogenetics in bringing validation to tree reconstruction methods. Experimental phylogenetics can offers

a true benchmark over computational benchmarks due to the added biological realism. Many argue that the reasons computational simulations are more popular is because they are cheaper and take an exponentially shorter amount of time to create compared to an experimental phylogeny. However, the primary limitation of numerical simulations is that they always include gross simplifications of biological processes. Being an advocate within the field of experimental phylogenetics however, does not negate the importance of computer simulations. On the contrary, the numerical simulation and experimental phylogeny approaches are largely complementary and both kinds of studies are essential for evaluating methods of phylogenetic analysis effectively.

## Creating an Experimental Phylogeny

Generating an experimental phylogeny in the laboratory requires a model organism of interest upon which to build. When considering a model organism, several important factors need to be considered. The organism should have short generation times, such as bacteria or phage in order to effectively evolve a data set large enough and diverse enough for phylogenetic analysis. Alternatively, gene sequences can be used in place of organisms and these sequences can be evolved in processes such as error prone PCR, where a generation time takes only a few hours dependant upon the length of the PCR cycle. Also, it is important to have a way to detect evolved mutants. Examples include antibiotic resistance in bacteria, plaque formation of phage, or functional assays of a

mutant protein. The phenotypic plasticity of the gene of interest should be considered, otherwise the experiment may fail to generate functional diversity.

## References


1. Vigilant, L., *et al.*, *Mitochondrial DNA sequences in single hairs from a southern African population.* Proceedings of the National Academy of Sciences, 1989. **86**(23): p. 9350-9354.
2. Chao, D.L., *et al.*, *The global spread of drug-resistant influenza.* Journal of The Royal Society Interface, 2012. **9**(69): p. 648-656.
3. Huelsenbeck, J.P. and D.M. Hillis, *Success of Phylogenetic Methods in the Four-Taxon Case.* Systematic Biology, 1993. **42**(3): p. 247-264.
4. Moret, B. and T. Warnow, *Reconstructing Optimal Phylogenetic Trees: A Challenge in Experimental Algorithmics.* Springer Berlin / Heidelberg, 2002. p. 163-180.
5. Felsenstein, J., *Evolutionary trees from DNA sequences: A maximum likelihood approach.* Journal of Molecular Evolution, 1981. **17**(6): p. 368-376.
6. Fitch, W.M., *Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology.* Syst Biol, 1971. **20**(4): p. 406-416.
7. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* Journal of Molecular Evolution, 1980. **16**(2): p. 111-120.

8. Rodriguez, F., et al., *The General Stocastic Model of Nucleotide Substitution.* J. ther. Biol., 1990. **142**: p. 485-501.

9. Felsenstein, J., *Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach.* J Mol Evol, 1981. **17**: p. 368-376.

10. Yang, Z., R. Nielsen, and M. Hasegawa, *Models of amino acid substitution and applications to mitochondrial protein evolution.* Mol Biol Evol, 1998. **15**(12): p. 1600-1611.

11. Yang, Z., et al., *Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites.* Genetics, 2000. **155**(1): p. 431-449.

12. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.

13. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21**(9): p. 2104-5.

14. Hillis, D.M., et al., *Experimental Phylogenetics: Generation of a Known Phylogeny.* Science, 1992. **255**(5044): p. 589-592.

15. Felsenstein, J., *Inferring Phylogeny*, ed. S. Associates2003, Sunderland, MA.

16. Sousa, A., et al., *Exploring tree-building methods and distinct molecular data to recover a known asymmetric phage phylogeny.* Molecular Phylogenetics and Evolution, 2008. **48**(2): p. 563-573.

17. Sanson, G.F.O., et al., *Experimental Phylogeny of Neutrally Evolving DNA Sequences Generated by a Bifurcate Series of Nested Polymerase Chain Reactions.* Mol Biol Evol, 2002. **19**(2): p. 170-178.

# CHAPTER 2

# ANCESTRAL SEQUENCE RECONSTRUCTION

## Introduction and Importance

In 1963, Emile Zuckerkandl and Linus Pauling published an article entitled "Molecular restoration studies of extinct forms of life" [1]. In it, they put forward the notion of reconstructing amino acid sequences of ancestral proteins by virtue of a comparison between sequences of related proteins found in contemporary organisms and subsequent synthesis (and thereby resurrection) of these sequences in the laboratory, and termed this 'paleogenetics'. While limits in technology prohibited the actual resurrection, Zuckerkandl and Pauling presented a sequence reconstruction of ancient mammalian hemoglobins. The duo then suggested that a future resurrection of ancient hemoglobins assayed for ancestral function (i.e. dioxygen affinity and pH dependence) would generate higher-order inferences of biological integration. Or, more specifically, the joining of chemical, biological, and structural models to natural history would provide a more accurate description of macromolecular behavior beyond that supplied by studying individual molecules disconnected from the selective forces governing their evolution.

The recent accumulation of DNA sequence data, combined with advances in evolutionary theory, computational power and DNA synthesis technology, have

paved the way for researchers to fulfill the vision of Zuckerkandl and Pauling [1-15]. Ancestral Sequence Reconstruction (ASR) allows us to infer ancestral gene sequences and then test the sequences in the laboratory by actually resurrecting ancient proteins themselves. Results from functional assays of the protein products from these ancient genes can then provide insight into their activities, interactions, binding-specificities, environments, etc.

Reconstruction of ancestral DNA and amino acid sequences is an important means of inferring information about past evolutionary events. ASR goes a step beyond that of phylogenetic reconstructions by determining the ancestral sequence's function. By resurrecting the ancestor and assaying its function, we can dissect evolutionary processes that have accrued over the course of evolutionary time. Phylogenetics lets us see how things connect, but ASR allows us to see how functions are lost or derived. To date, approximately 20 narratives have emerged where specific molecular systems from extinct organisms have been resurrected for study in the laboratory [2-23]. These systems include digestive proteins in ruminants and primates to illustrate how digestive function arose from non-digestive function, fermentive enzymes from fungi, pigments in the visual system adapting to different environments, steroid hormone receptors adapting to changing function in steroid-based regulation of metazoans, fluorescent proteins from ocean-dwelling invertebrates, enzyme co-factor evolution, proteins from very ancient bacteria helping to define environments where early forms of life lived, *inter alia.*

## Methods

ASR uses a present-day backwards strategy, whereby you create a multiple sequence alignment using gene sequences of interest, construct a phylogeny from the alignment, infer the ancestral sequences located at the nodes of the tree, and lastly, resurrect the sequences in the laboratory via gene synthesis and recombinant expression in order to assay the ancient protein's function. Similar to constructing a phylogenetic tree, ancestral sequences may be inferred from a variety of computational methods including maximum parsimony, maximum likelihood, and Bayesian methods. Although parsimony methods were popular among the first ancestral reconstructions, most modern studies use maximum likelihood (ML) [24]. ML and Bayesian methods are similar in that they calculate ancestral sequences using a tree topology and evolutionary model, however the ML method will always create the most likely ancestral sequence or the 'most probable ancestral sequence (MPAS) while Bayesian methods incorporate uncertainties associated with the tree and/or model when reconstructing the ancestral sequence [25].

## Criticism

ASR is mostly limited by its inability to know the 'truth' of ancestral character states because neither the sequence nor the behavior of a protein from an organism that went extinct a billion years ago can be known with the same precision as the sequence or behavior of a descendent living today [26]. ASR practitioners acknowledge a certain degree of inaccuracy associated with

reconstructed ancestral sequences, however, developments in ancestral reconstruction methods allow for the accurate reconstruction of more ancient proteins than previously thought possible [27].

ASR accuracy has been determined by computer simulations. Within these simulations, sequences evolve under evolutionary models according to a given tree topology. The tip sequences of the simulated phylogenies are then used in ASR analyses, and ancestral sequence inferences are compared with the 'known' ancestral sequences generated by the simulation. In the early incarnations, the simulations involved simple, non-biologically relevant conditions such as the classic 4-taxon tree in which sequences were simulated across the tree and then the simulated sequences were used to infer the ancestral states from which they were simulated. In the most recent incarnation, Prof. Thornton's group has computationally simulated sequences across biologically relevant phylogenies [28]. Here, the Thornton group used topologies and branch-lengths from multiple ASR studies to serve as the phylogenies for the simulations. Sequences were then simulated across these phylogenies and the simulated sequences were used to determine the differences in performance between empirical and hierarchical Bayesian approaches for inferring ancestral character states.

Despite simulations serving as good first approximations for the performance and accuracy of ancestral sequence reconstruction (ASR), they fail to capture the fundamental aspect of all ASR studies – functional/phenotypic

diversification. All gene families and phylogenies used in ASR studies contain examples of functional diversification among descendent sequences (e.g., ligand binding, fluorescent color emission, co-factor binding, protein thermostability, etc.), which is reasonable because there would be little value of performing ASR on a gene family when all the descendent sequences have the exact same properties and behaviors. It is thus very important to benchmark ASR methods against datasets that contain functional divergence of some phenotype or functional property.

Functional divergence of biological sequences is often captured when analyzing sequences using a nonsynonymous-to-synonymous ratio (branch-specific or sites-specific) and/or other metrics such as Type-I functional divergence (e.g., covarion, heterotachy, etc.) or Type-II functional divergence [29, 30]. It would thus be valuable if a simulation study had, say, branches with high nonsynonymous-to-synonymous ratios in which a burst of amino acid replacements represented a change in phenotype during the simulation. This is of course difficult to simulate even with a computer lattice model because we do not know precisely how biological proteins evolve modified functions and behaviors due to a gap in our precise understanding of the connection between genotype and phenotype. Along these lines, it has been noted that experimental approaches may be advantageous to computational simulations particularly when biological reality cannot be explicitly modeled [31]. Naylor and colleagues

precisely supported this notion recently by demonstrating how simulations may fail to specify biological reality [32].

Again, computer simulations of ancestral genotypes and phenotypes have offered an intriguing first-approximation for reality, however, in order to expand the recent success of ASR and add quantitative rigor to the field a more appropriate benchmark of method performance requires an evaluation of biological sequences and phenotypes measured in the laboratory. The work presented in this thesis attempts to provide such a benchmark.

## References


1. Pauling, L. and E. Zuckerkandl, *Chemical Paleogenetics: Molecular Resoration Studies of Extinct Forms of Life.* Acta Chemica Scandinavica, 1963. **17**: p. 9-16.
2. Adey, N.B., et al., *Molecular Resurrection Of An Extinct Ancestral Promoter For Mouse L1.* Proceedings Of The National Academy Of Sciences Of The United States Of America, 1994. **91**(4): p. 1569-1573.
3. Benner, S.A., *Reconstructing Ancient Forms Of Life.* Journal Of Cellular Biochemistry, 1995: p. 200-200.
4. Blanchette, M., et al., *Reconstructing large regions of an ancestral mammalian genome in silico.* Genome Res, 2004. **14**(12): p. 2412-23.
5. Bridgham, J.T., S.M. Carroll, and J.W. Thornton, *Evolution of hormone-receptor complexity by molecular exploitation.* Science, 2006. **312**(5770): p. 97-101.

6. Chandrasekharan, U.M., et al., *Angiotensin II-forming activity in a reconstructed ancestral chymase.* Science, 1996. **271**(5248): p. 502-5.

7. Chang, B.S., et al., *Recreating a functional ancestral archosaur visual pigment.* Mol Biol Evol, 2002. **19**(9): p. 1483-9.

8. Galtier, N., N. Tourasse, and M. Gouy, *A nonhyperthermophilic common ancestor to extant life forms.* Science, 1999. **283**(5399): p. 220-221.

9. Gaucher, E.A., et al., *Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins.* Nature, 2003. **425**(6955): p. 285-8.

10. Ivics, Z., *et al*., *Molecular reconstruction of Sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells.* Cell, 1997. **91**(4): p. 501-510.

11. Jermann, T.M., *et al*., *Reconstructing The Evolutionary History Of The Artiodactyl Ribonuclease Superfamily.* Nature, 1995. **374**(6517): p. 57-59.

12. Krishnan, N.M., *et al*., *Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference.* Molecular Biology And Evolution, 2004. **21**(10): p. 1871-1883.

13. Malcolm, B.A., *et al*., *Ancestral Lysozymes Reconstructed, Neutrality Tested, And Thermostability Linked To Hydrocarbon Packing.* Nature, 1990. **345**(6270): p. 86-89.

14. Messier, W. and C.B. Stewart, *Episodic adaptive evolution of primate lysozymes.* Nature, 1997. **385**: p. 151 - 154.

15. Stackhouse, J., *et al.*, *The Ribonuclease From An Extinct Bovid Ruminant.* Febs Letters, 1990. **262**(1): p. 104-106.

16. Thomson, J.M., *et al.*, *Resurrecting ancestral alcohol dehydrogenases from yeast.* Nat Genet, 2005.

17. Thornton, J.W., E. Need, and D. Crews, *Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling.* Science, 2003. **301**(5640): p. 1714-7.

18. Ugalde, J.A., B.S. Chang, and M.V. Matz, *Evolution of coral pigments recreated.* Science, 2004. **305**(5689): p. 1433.

19. Zhang, J., *Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys.* Nat Genet, 2006. **38**(7): p. 819-23.

20. Zhang, J.Z. and M. Nei, *Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods.* Journal Of Molecular Evolution, 1997. **44**: p. S139-S146.

21. Zhu, G.P., G.B. Golding, and A.M. Dean, *The selective cause of an ancient adaptation.* Science, 2005. **307**(5713): p. 1279-1282.

22. Ortlund, E.A., *et al.*, *Crystal structure of an ancient protein: Evolution by conformational epistasis.* Science, 2007. **317**(5844): p. 1544-1548.

23. Gaucher, E.A., S. Govindarajan, and O.K. Ganesh, *Palaeotemperature trend for Precambrian life inferred from resurrected proteins.* Nature, 2008. **451**(7179): p. 704-7.

24. Yang, Z., S. Kumar, and M. Nei, *A New Method of Inference of Ancestral Nucleotide andAmino Acid Sequences.* Genetics, 1995. **141**: p. 1641-1650.

25. Pagel, M., A. Meade, and D. Barker, *Bayesian estimation of ancestral character states on phylogenies.* Systematic Biology, 2004. **53**(5): p. 673-84.

26. Williams, P.D., *et al.*, *Assessing the Accuracy of Ancestral Protein Reconstruction Methods.* PLoS Comput Biol, 2006. **2**(6): p. e69.

27. Chang, B.S.W., J.A. Ugalde, and M.V. Matz, *Applications of Ancestral Protein Reconstruction in Understanding Protein Function: GFP-Like Proteins*, in *Methods in Enzymology*, A.Z. Elizabeth and H.R. Eric, Editors. 2005, Academic Press. p. 652-670.

28. Hanson-Smith, V., B. Kolaczkowski, and J.W. Thornton, *Robustness of ancestral sequence reconstruction to phylogenetic uncertainty.* Molecular Biology and Evolution, 2010. **27**(9): p. 1988-99.

29. Gaucher, E.A., *et al.*, *Predicting functional divergence in protein evolution by site-specific rate shifts.* Trends In Biochemical Sciences, 2002. **27**(6): p. 315-321.

30. Gaucher, E.A., M.M. Miyamoto, and S.A. Benner, *Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein.* Genetics, 2003. **163**: p. 1549 - 1553.

31. Oakley, T.H., *A critique of experimental phylogenetics*, in *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*, T. Garland Jr. and M.R. Rose, Editors. 2009, University of California Press. p. 659-669.

32. Lakner, C., *et al*., *What's in a Likelihood? Simple Models of Protein Evolution and the Contribution of Structurally Viable Reconstructions to the Likelihood.* Systematic Biology, 2011. **60**(2): p. 161-174.

# CHAPTER 3

# EXPERIMENTAL PHYLOGENETICS: A BENCHMARK FOR ANCESTRAL SEQUENCE RECONSTRUCTION

To overcome potential barriers associated with benchmarking ASR using computer simulations, we have generated a laboratory phylogeny derived from the accelerated evolution and artificial selection of fluorescent protein variants. Though the experimental phylogeny is not yet complete, we have begun to create a system for generating a biologically relevant phylogeny containing biologically relevant sequences that experienced functional diversification (unlike simulations). When the experimental phylogeny is complete, its tip sequences will then be used to benchmark various ASR approaches under various phylogenetic conditions.

As such, we have used members of the fluorescent protein family to generate an experimental phylogeny, à la Hillis & Bull [1-8]. The fluorescent protein family is widely used to study in vitro directed evolution techniques due, in part, to the simplicity of phenotypic assays, wide range of emission spectra requiring a relativity modest amount of mutation, and small sequence length (~230 amino acids) [9-11]. These properties, by extension, make the family an ideal choice for experimental phylogenetic studies.

## Fluorescent Proteins

Members of the fluorescent protein (FP) family have a unique cylindrical molecular structure. Their peptide sequence is organized into an eleven-stranded β-sheet, a major a-helix, and small a-helices connecting the ends of the anti-parallel β-strands [12-15]. The protein is in the shape of a cylinder, comprising a β-barrel with a single a-helix running through the inside and short a-helical segments capping the ends of the β-barrel. This motif results in a very uniform and tightly compacted β-can of about 30 Å in diameter and 40 Å in length (Figure 3.1). FP's are able to fluorescence through formation of an intrinsic chromophore. The chromophore is situated within the geometric center of the β-barrel and arises from the covalent modification of three adjacent amino acids in the folded FP structure. Fluorescence occurs when the chromophore absorbs visible light, donates an electron in its excited state, and emits visible light at a longer wavelength. Color arises when the chromophore absorbs certain wavelengths of visible light and transmits others.

Our experimental phylogeny is derived from the monomeric red fluorescent protein (mRFP 1.0), an engineered monomeric version of the natural tetrameric red fluorescent protein isolated from the coral species *Disconsoma straita* (DsRed) [16, 17]. The first fluorescent protein ever isolated was the green fluorescent protein (GFP) from the Jellyfish *Aquorea victoria* and until the discovery of DsRed in 2001 within the nonbiolumenescent Anthozoa coral species, all fluorescent proteins were thought to be green [18, 19]. Coral FPs,

however, are renowned for their functional divergence/color variety and are classified into five main spectral groups: cyan (CFPs), green (GFPs), red (RFPs), and non-fluorescent chromoproteins (CPs) [20] (Table 3.1). Other spectral FP variants include blue and orange derivatives engineered for use in cell imaging assays [21-23].

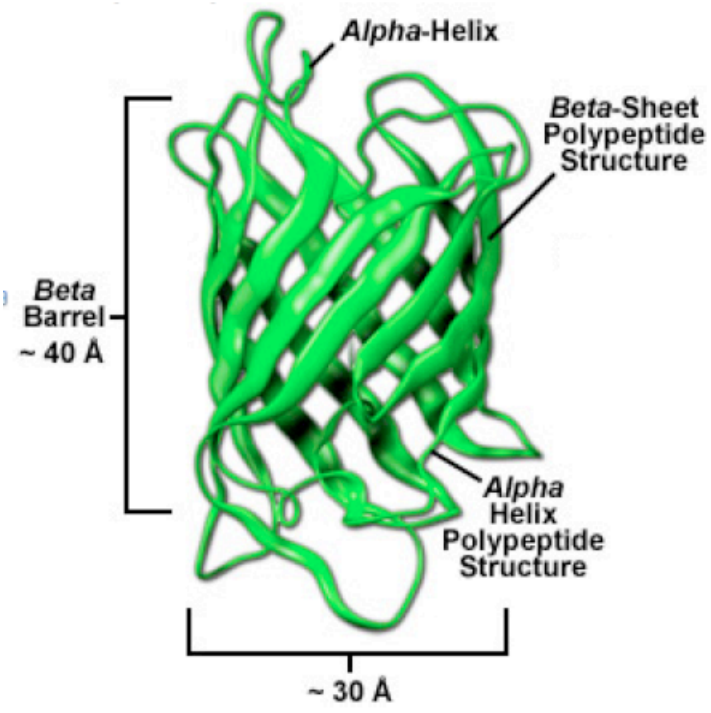



**Figure 3.1. Fluorescent Protein Structure.** Fluorescent protein β-barrel architecture and approximate dimensions. Image taken from [12].

**Table 3.1. Excitation and emission wavelengths of fluorescent proteins found within coral species.** Four types of proteins within the fluorescent protein superfamily have been identified in corals: cyan, green, red, and a non-fluorescent purple chromoprotein. Excitation and emission maxima of each protein are listed to the right of the protein.

| Coral Color Classes | max excitation nm | max emission nm |
|---|---|---|
| Cyan | 404-467 | 485-495 |
| Green | 478-512 | 500-524 |
| Red | 560-578 | 576-595 |
| Chromoprotein | 560-588 | not fluorescent |

## Evolution of Fluorescent Proteins

In attempts to understand FP color diversity, Shagin *et al*. assessed the deep level phylogenetic relationships within the FP superfamily and found that the origin of red, yellow, and cyan fluorescence occurred independently on several occasions, providing a remarkable example of convergent evolution with complex features at the molecular level [24]. Further phylogenetic analysis supports various episodes of convergent evolution and further suggests that color diversity has originated recently and independently within several lineages [11, 25].

Also, recent research shows that the mutations responsible for the generation of modern multi-colored FP phenotypes among corals from the green FP ancestor [26] are shown to have arisen by mutations that were fixed due to positive natural selection [27]. Evolution by such a mechanism would explain the recent diversification of colors and suggest that coral FP color diversity is a product of adaptive evolution [28].

## Aim

In light with the added biological realism within experimental phylogenies, our experimental phylogeny was built in accordance with natural FP evolution in a way to recapitulate the adaptive radiation events within the natural FP phylogeny. Given the rapid evolution of FP colors, the mRFP 1.0 red fluorescent 'ancestor' experiences punctuated equilibrium, with bursts of evolutionary change and rapid events of branching speciation then followed then by an extended state or stasis, whereby the species under go little evolutionary change over time.

The research described in this thesis attempts to add rigor to the field of ASR by generating a known and phenotypically-diverse phylogeny in the laboratory based on random mutation and artificial selection of fluorescent proteins. The evolved tip sequences on the experimental phylogeny will then be used to resurrect ancient sequences and the inferred ancestral sequences will be compared to the true ancestral sequences within the experimental phylogeny. In doing so, we will for the first time be able to experimentally validate and benchmark ASR.

## References


1. Bull, J.J., *et al.*, *Experimental evolution yields hundreds of mutations in a functional viral genome.* Journal Of Molecular Evolution, 2003. **57**(3): p. 241-248.
2. Bull, J.J., *et al.*, *Genome properties and the limits of adaptation in bacteriophages.* Evolution, 2004. **58**(4): p. 692-701.
3. Bull, J.J., *et al.*, *Exceptional convergent evolution in a virus.* Genetics, 1997. **147**(4): p. 1497-1507.
4. Bull, J.J., *et al.*, *Experimental Molecular Evolution of Bacteriophage-T7.* Evolution, 1993. **47**(4): p. 993-1007.
5. Hillis, D.M., *Approaches For Assessing Phylogenetic Accuracy.* Systematic Biology, 1995. **44**(1): p. 3-16.
6. Hillis, D.M., *et al.*, *Experimental Phylogenetics - Generation of a Known Phylogeny.* Science, 1992. **255**(5044): p. 589-592.
7. Hillis, D.M., *et al.*, *Experimental Approaches to Phylogenetic Analysis.* Systematic Biology, 1993. **42**(1): p. 90-92.
8. Sanson, G.F.O., *et al.*, *Experimental Phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions.* Molecular Biology And Evolution, 2002. **19**(2): p. 170-178.
9. Campbell, R.E., *et al.*, *A monomeric red fluorescent protein.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7877-82.

10. Shaner, N.C., *et al.*, *Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein.* Nat Biotechnol, 2004. **22**(12): p. 1567-72.

11. Kelmanson, I.V. and M.V. Matz, *Molecular basis and evolutionary origins of color diversity in great star coral Montastraea cavernosa (Scleractinia: Faviida).* Molecular Biology and Evolution, 2003. **20**(7): p. 1125-33.

12. Sample, V., R.H. Newman, and J. Zhang, *The structure and function of fluorescent proteins.* Chemical Society reviews, 2009. **38**(10): p. 2852-64.

13. Shaner, N.C., G.H. Patterson, and M.W. Davidson, *Advances in fluorescent protein technology.* Journal of cell science, 2007. **120**(Pt 24): p. 4247-60.

14. Zimmer, M., *Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior.* Chem. Rev., 2002. **102**: p. 759-781.

15. Ormo, M., *et al.*, *Crystal Structure of the Aequorea victoria Green Fluorescent Protein.* Science, 1996. **273**(5280): p. 1392-1395.

16. Campbell, R.E., *et al.*, *A monomeric red fluorescent protein.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(12): p. 7877-82.

17. Yarbrough, D., *et al.*, *Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-Å resolution.* PNAS, 2001. **98**(2): p. 462-467.

18. Tsien, R.Y., *THE GREEN FLUORESCENT PROTEIN.* Annual Review of Biochemistry, 1998. **67**: p. 590-44.

19. Alieva, N.O., *et al.*, *Diversity and evolution of coral fluorescent proteins.* PLoS ONE, 2008. **3**(7): p. e2680.

20. Hunt, M.E., *et al.*, *Multi-domain GFP-like proteins from two species of marine hydrozoans.* Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology, 2012. **11**(4): p. 637-44.

21. Zhang, J., *et al.*, *Creating new fluorescent probes for cell biology.* Nature reviews. Molecular cell biology, 2002. **3**(12): p. 906-18.

22. Remington, S.J., *Green fluorescent protein: a perspective.* Protein science : a publication of the Protein Society, 2011. **20**(9): p. 1509-19.

23. Rizzo, M.A., M.W. Davidson, and D.W. Piston, *Fluorescent Protein Tracking and Detection: Fluorescent Protein Structure and Color Variants.* Cold Spring Harbor Protocols, 2009. **2009**(12): p. pdb.top63.

24. Shagin, D.A., *et al.*, *GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity.* Molecular Biology and Evolution, 2004. **21**(5): p. 841-50.

25. Labas, Y.A., *et al.*, *Diversity and evolution of the green fluorescent protein family.* Proceedings of the National Academy of Sciences, 2002. **99**(7): p. 4256-4261.

26. Ugalde, J.A., B.S.W. Chang, and M.V. Matz, *Evolution of Coral Pigments Recreated.* Science, 2004. **305**(5689): p. 1433.

27. Wachter, R.M., J.L. Watkins, and H. Kim, *Mechanistic Diversity of Red Fluorescence Acquisition by GFP-like Proteins.* Biochemistry, 2010. **49**(35): p. 7417-7427.

28. Field, S., *et al*., *Adaptive Evolution of Multicolored Fluorescent Proteins in Reef-Building Corals.* Journal of Molecular Evolution, 2006. **62**(3): p. 332-339.

# CHAPTER 4

# MATERIALS AND METHODS

## Materials

### *Escherichia coli* strains and genotypes

NovaBlue Competent Cells (Novagen, Gibbstown, NJ):

Chemically competent *E. coli* K-12 strain used for plasmid transformation and provides high yields of plasmid DNA. Genotype: *endA1 hsdR17* ($r_{k12}$-$m_{k12}^{+}$) *supE44 thi-1 recA1 gyrA96 relA1 lac* F′[*proA*$^{+}$*B*$^{+}$ *lacI*$^{q}$ZΔM15::Tn*10*] (Tet$^{R}$).

BL21(DE3) Competent Cells (Novagen):

Chemically competent *E. coli* B strain used for plasmid transformation and protein expression. These cells carry a chromosomal copy of the T7 RNA polymerase gene under control of the lacUV5 promoter whereby target genes cloned into pET vectors are expressed upon IPTG induction. Genotype: F$^{-}$ *ompT hsdS*B ($r_B^{-}$ $m_B^{-}$) *gal dcm* (DE3).

### Vector

pET-15b (Novagen):

This vector contains a T7 RNA polymerase promoter and a N-terminal His-Tag followed by a thrombin cleavage site. This vector was used for cloning and expression of fluorescent proteins. Antibiotic resistance: Carbenicilin (CARB).

**General Media and Buffers**

Carbenicillin (CARB) Stock:

For 100 mg/mL: 1 g of CARB was added to 9 mL of DI water, filter sterilized and stored at -20 °C.

Isopropyl β-D-1-thiogalactopyranoside (IPTG) Stock:

281.3 mg of IPTG was added to 9.76 mL of DI water, filter sterilized, and stored at -20 °C.

LB (Luria-Bertani) Media:

0.5 % yeast extract, 1 % tryptone, 0.5 % sodium chloride (NaCl), and 1.6 % agar for plates. For media containing CARB, 1 mL of 100 mg/mL CARB was added to 1 L autoclaved media (final 100 μL/mL). For media containing IPTG, 200 μL of 100 mM IPTG was added to 1 L autoclaved media (final 20 μM).

YETM:

0.5 % yeast extract, 2 % tryptone, 1 % magnesium sulfate heptahydrate ($MgSO_4 \cdot 7H_2O$), and 1.5% agar for plates. Adjust to pH 7.5 with KOH.

TFB1:

30 mM potassium acetate (KOAc), 100 mM Rubidium chloride (RbCl), 10 mM $CaCl_2 \cdot 2H_2O$, 50 mM $MnCl_2 \cdot 4H_2O$, and 15% Glycerol. Adjust to pH 5.8 with 0.2 M acetic acid. Store at 4 °C.

TFB2:

10 mM MOPS, 75 mM $CaCl_2$-$2H_2O$, 10 mM RbCl, and 15 % Glycerol. Adjust to pH 6.5 with KOH. Store at 4 °C.

10X TBE Buffer:

10.8 % Tris base, 5.5 % Boric acid, and 0.93 % EDTA

TBE agarose gel:

1X TBE buffer, 0.8 – 1% agarose, and 0.5% EtBr.

### Genes and Primers

mRFP Gene:

The monomeric red fluorescent protein (mRFP) is the gene used in this study. mRFP was donated by A. Bommarius (Georgia Tech).

3s.1 ASR Genes:

Ancestral sequence reconstructions (ASR) of 3s.1 sequences were synthesized and cloned into pET-15b (GenScript, Piscataway, NJ). Genes were diluted to 40 ng/μL with sterile DI water and were stored at -20 °C.

Primers:

Synthetic PCR primers were purchased from Integrated DNA Technologies (IDT) Inc (San Diego, CA). Primers were used for subcloning and mutagenesis reactions. Primers were diluted to 100 μM with sterile DI water and stored at -20 °C.

## Methods

### Preparation of Competent Cells

The purchased frozen cells were streaked onto a 100 X 15 mm petri dish containing YETM agar media and incubated at 37 °C. A single colony from YETM

plate was inoculated into 5 mL of YETM medium and incubated overnight. The overnight culture was diluted into 250 mL YETM medium and incubated until the culture reached an optical density ($OD_{600}$) of 0.4 – 0.8. 50 mL of the culture was aliquoted into 5 centrifuge tubes. The rest of the procedure was performed as follows: cultures were incubated on ice for 10 min, centrifuged for 10 min at 4 °C 2000 rpm, supernatant discarded, pellets were resuspended in 10 ml of TFB1 and incubated on ice for 5 min, centrifuged for 10 min at 4 °C 2000 rpm, supernatant discarded, pellet were resuspended in 2 mL of TFB2 and incubated on ice for 15 min. The competent cells were stored at -80 °C in 60 - 210 μL aliquots. Transformation of the cells was performed with using the mRFP 1.0 construct and plated on LB, IPTG, CARB plate to assess contamination and transformation efficiency.

**Freezer Stocks**

A mix of 100 μL dimethyl sulfoxide (DMSO) and 900 μL cells containing recombinant plasmids were stored at -80 °C.

**Plasmid purification**

All plasmids were purified from *E. coli* cells using Qiagen's (Valencia, CA) QIAprep Spin Miniprep Kit following the manufacture's protocol with the following exception: DNA was eluted in 30 μL of water.

**DNA Quantification**

DNA concentrations were determined using a NanoDrop 1000

Spectrophotometer version 3.7 (Thermo Scientific).

**PCR and Digest clean up**

PCR and digest reactions were purified using Qiagen's QIAquick PCR Purification Kit following the manufacture's protocol with the following exception: DNA was eluted in 30 μL of water.

**Digestion**

All digests use restriction enzymes *NdeI* (NEB) and *XhoI* (NEB): 5' (CATATG) and 3' (CTCGAG). Digestions were performed in 20 - 100 μL that include DNA, *NdeI*, *XhoI*, Buffer 4 (NEB) and Bovine serum albumin (BSA) (NEB). Digests were set up and incubated according to the manufacture's protocol.

**Ligation**

Purified digests were ligated in 10 - 30 μL reactions at room temperature (temp) for 3 hrs containing T4 Ligase (New England Biolabs (NEB), Ipswich, MA) and T4 Ligase buffer (NEB) set up according to manufactures protocol. mRFP genes were ligated into pET-15b in a molecular weight ratio of 1:4.

**Chemically Competent Transformation**

All transformations were performed as follows: 1 – 5 μL DNA was swirled into cells on ice and was incubated on ice 10 min, heat shock 30 sec 42 °C, incubated on ice 2 min. 250 μL of LB was added to the transformation, and cells recovered for 30 min-1.5hr in a 37 °C incubator shaking at 250 rpm.

Mutagenesis transformations:

3-5 uL of ligation was swirled into 20 μL BL21(DE3) cells on ice. Two to eight transformations were spread on 22.5 cm square plates containing LB, CARB, IPTG and then incubated at 37 °C for 12 - 20 hrs.

Plasmid Transformations:

25 ng of DNA was swirled into 20 μL BL21(DE3) or 10 μL of NovaBlue cells on ice. 25 -50 μL of transformation was spread on 100 X 15 mm petri dish containing LB, CARB, IPTG (or without IPTG if protein expression was not necessary or if using NovaBlue cells) and then incubated at 37 °C for 12 - 20 hrs.

**Subcloning mRFP into pET-15b**

DH5a *E. coli* cells containing mRFP gene (678 bp) cloned in the pPROTet.E vector (Clontech, Mountain View, CA) were obtained from A. Bommarius (Georgia Tech). The using the QIAprep Spin Miniprep Kit. The following primers were used for subcloning mRFP from pPROTet.E to expression vector pET-15b : mRFP-5'-NdeI (TATTCATATGGCGTCTTCTGAAGACGTTATC) and mRFP-3'-XhoI (TATTCTCGAGCTATTACGCACCGGTAGAGTG). The PCR reaction consisted of the following: 30 ng purified plasmid, 0.25 μL of Phusion® High-Fidelity DNA Polymerase (NEB), 5 μL of 5X Phusion® HF Buffer (NEB), 0.5 μL of 100 mM dNTPs (Promega, Madison, WI), 2.5 μL 10 μM mRFP-5'-NdeI, and 2.5 μL 10 μM mRFP-3'XhoI brought up to a final volume of 25 μL with water. The reaction ran under the following cycling parameters: initial incubation at 98 °C 30 sec then 98 °C 10 sec, 59 °C 30 sec, 72 °C 15 sec, repeat X24, final incubation at 72 °C 5 min. 15 μL mRFP amplified products were digested using restriction

enzymes *NdeI* (NEB) and *XhoI* (NEB): 5' (CATATG) and 3' (CTCGAG). The digestion was performed in a 20 μL reaction that included 6 μg PCR product, 1 μL *NdeI*, 1 μL *XhoI*, 2 μL Buffer 4 (NEB) and was incubated in a 37°C water bath for 6 hours (hrs). *NdeI* and *XhoI* digestion pET-15b was performed under the same conditions as described above except for the following: 4 μg pET-15b was used in place of the 6 μg mRFP PCR product. Digest was cleaned up using the QIAquick PCR Purification Kit and ligated in a 25 μL reaction containing the following: 20 ng insert (digested mRFP PCR product), 100 ng digested pET-15b vector, 1 μL T4 Ligase (Promega), 2.5 μL Ligase Buffer (Promega). 5 μL ligation reaction was transformed into 25 μL NovaBlue Competent Cells (Novagen). 50 μL transformation was spread onto a 100 X 15 mm petri dish containing LB, CARB, agar was incubated overnight at 37 °C.

Selection

The next day, 8 colonies from the transformation plate were inoculated individually into 5 mL LB containing 100 μg/mL carb and were incubated in a 37 °C incubator shaking at 250 rpm overnight. The overnight cultures were purified using Qiagen's QIAprep Spin Miniprep Kit following the manufacture's protocol and all samples yielded ~550 ng of purified plasmid DNA. 5 μL of purified samples were digested in 1 μL *XhoI*, 1 μL *NdeI*, 2 μL Buffer 4, and deionized (DI) water up to a total volume of 20 μL; digestions sat in a 37°C water bath overnight. In a 1.0% TBE-EtBr gel, 10 μL digested samples + 2 μL 10X loading dye were loaded into individual lanes and 10 μL 1 Kb Ladder (Fisher Scientific)

was loaded into its own lane. DNA was visualized within a UV box (REF). The DNA sample selected showed two bands: one band at about 6,000 bp (pET-15b Vector) and the other band at about 700 bp (mRFP gene). 480 ng of the mRFP-pET-15b construct was sent to GeneWiz (South Plainfield, NJ) for sequencing with T7 and T7-Rev universal primers. Sequence results were analyzed using CLC Bio (Cambridge, Massachusetts) software v.4.1.2, and the pET-15b/mRFP construct was confirmed; this construct was designated mRFP 1.0.

**Building the Experimental Phylogeny**

Primers (IDT) were synthesized for use in error prone PCR:

mRFP Random For (GGCAGCCATATGGCGTCTTCTGAAGACGTTATC)

mRFP Random Rev (CGGATCCTCGAGCTATTACGCACCGGTAGAGTG)

Primer mRFP Random For was later replaced by primer FWD mRFP NdeI (CTGGTCGGCCATATGGCGTCTTCTGAAGACGTTATC) since the latter primer increases the number of bases flanking *NdeI*'s recognition sequence and thereby increasing restriction efficiency. Random mutagenesis of mRFP 1.0 was performed using the GeneMorph II Random Mutagenesis Kit (Stratagene, La Jolla, CA). Each reaction was performed in 50 μL and consisted of the following: 425 – 625 ng template plasmid, 0.25 μL forward primer, 0.25 μL reverse primer, 1μL of 40 mM dNTP stock, 5 μL 10X Mutazyme II reaction buffer, 1 μL Mutazyme II DNA polymerase. PCR was performed using the following conditions: initial incubation at 95 °C 2 min then 95° C 30 sec, 59° C 30 sec, 72° C 1 min, repeat x29, final incubation at 72° C 10 min. PCR products were purified using Qiagen's

PCR clean up kit following the manufacture's protocol, except for the following: DNA was eluted in 30 - 46.5 μL water. Purified mRFP 1.0 mutagenesis reaction was digested in a 50 μL reaction at 37 °C from 16 - 48 hrs and included the following: pure mRFP 1.0 mutants, 1 μL *XhoI*, 1 μL *NdeI*, 5 μL Buffer 4, 0.5 μL BSA. pET-15B digetsions cut with *NdeI* and *XhoI* included 1.5 μg of pET-15b vector, 1 μL *XhoI*, 1 μL *NdeI*, 5 μL Buffer 4, 0.5 μL BSA, and DI water up to a total volume of 50 μL; usually 10 of these digestions were set up at a time. mRFP mutant digest was purified using Qiagen's PCR clean up kit following the manufacture's protocol, except for the following: DNA was eluted in 30 μL water. The 10 pET-15b *XhoI* and *NdeI* digests were combined and were purified using Qiagen's PCR clean up kit following the manufacture's protocol, except for the following: DNA was eluted in 30 μL water. The concentrations of the purified digests were quantified using a nanospec (REF). Fluorescent protein genes were ligated into pET-15b according to the following protocol: 100 ng digested pET-15b Vector, 20 ng digested FP gene, 0.5 μL T4 Ligase, 1 - 2 μL 10X T4 Ligase Buffer in a 10 – 20 μL reaction. Plasmids containing mutated mRFP 1.0 were transformed into expression host *E. coli* BL21(DE3).

Selection

Eight to twenty colonies expressing either wild-type fluorescent color or a mutant fluorescent color as determined using a hand-held ultraviolet lamp emitting 365 nm light were picked and inoculated into 3 - 5 mL of LB, Carb and grown overnight at 37 °C. Minipreps were performed the following day and submitted for

sequencing (GeneWiz). Sequence data were analyzed using CLC Bio (Cambridge, Massachusetts) software version 4.1.2. The average variant contained 1 - 4 mutations per round of random mutagenesis given the conditions above (after many trials to optimize the mutation load). One mutant was retained after each round of mutagenesis and used for subsequent rounds of random mutagenesis. Mutants were selected to balance the frequency of synonymous and nonsynonymous mutations along branches of the experimental gene phylogeny. In some instances two mutants were retained after a round of mutagenesis in lieu of one due to a speciation event.

**FP Spectra**

Fluorescent colonies were visualized and their spectra analyzed using a Zeiss LSM 510 NLO META confocal system with motorized stage and with laser lines 453 nm, 488 nm, and 514nm provided by an argon12 laser, and laser line 543 nm provided by HeNe1 HeNe1 laser. We used the Plan-Neofluar 100X/1.3 oil lens and the HAL100 camera, along with Zen2009 software. Scans were performed in lambda mode and both rhodamine and FITC filters were used to visualize red-type and green-type fluorescence, respectively.

**Experimental Phylogeny Analysis**

Finding the Best-Fit Evolutionary Models:

jModelTest version 0.1.1 was applied to the 19 terminal DNA sequences and ProtTest version 2.4 was applied to the 19 encoded terminal amino acid sequences in order to determine which evolutionary models and parameters best

fit our data using hierarchical Likelihood Ratio Tests (hLRT) and Akaike's Information Criterion (AIC) [1-3]. The best-fit nucleotide substitution model was GTR+I and the best-fit amino acid substitution model was HIVb+I+G.

Parameter Estimates of the Experimental FP tree and Natural FP Genes

19 randomly selected proteins from the green fluorescent protein (GFP) family were chosen and aligned (GFP19). PAML version 4.5 codeml ran with the 19 terminal codon sequences from both the 19 tips of the mRFP phylogeny and GFP19; the codon alignment is based on proteins using PAL2NAL version 14. Parameters of mRFP were estimated using three different tree topologies (topologies are explained in detail in section below titled "Phylogenetic Tree Inferences" in Chapter 6) were used, true tree topology using DNA input (TT_DNA), true tree topology using amino acid input (TT_AA), MrBayes version 3.2.1 DNA tree topology (MB_DNA), and MrBayes amino acid tree topology (MB_AA). All three give similar parameters. Parameters of GFP19 were estimated using 1) DNA and 2) amino acid tree topologies inferred from MrBayes and ran for 2 million generations. Both topologies give similar outputs. The GFP19 MrBayes 2 million DNA tree topology and the mRFP MrBayes 20 million DNA tree topology were used in conjunction with GFP19 terminal sequences and mRFP 19 terminal sequences, respectively, for comparative analysis of parameter estimates.

**Ancestral Sequence Reconstruction**

Sequences at the tips (leaves) of the mRFP evolved phylogeny are used to

computationally reconstruct the inferred ancestral fluorescent sequences at node 3.s1 of the tree. DNA and amino acid-based approaches are exploited, as well as MP and various ML methods.

Phylogenetic Tree Inferences:

The nineteen terminal DNA sequences ran for ten and twenty million generations in MrBayes under the best-fit GTR+I model as predicted by jModeltest 0.1.1 via AIC model selection. Both trees generated the exact same tree topology with nearly identical posterior probabilities (PP) for branch lengths. The second best-fit DNA model by jModelTest via AIC model selection is the GTR+I+G model. The nineteen terminal DNA sequences ran for ten million generations in MrBayes under the GTR+I+G model and resulted in the exact same tree topology and similar PP compared to the DNA tree ran under the GTR+I model. The nineteen terminal amino acid sequences ran for four and ten million generations in MrBayes under the JTT+I+G model predicted by ProtTest 2.4 Mac version via AIC model selection (JTT+I+G). Both trees generated the same tree topology and similar branch length P. The true tree topology of the evolved laboratory FPs was constructed using PAML. The GTR+I DNA topology (MB_DNA) and the JTT+I+G amino acid topology (MB_AA) were selected for were chosen for use in ASR analysis.

Sequence Reconstruction:

Maximum parsimony (MP) and Maximum likelihood (ML) inferences of 3s.1 were performed. To test for discrepancies among the models of evolution used for

sequence reconstruction in the ML method, multiple models were chosen. For DNA topologies (MB_DNA and TT_DNA), sequences were inferred under the GTR+G (JModelTest Rank #13), GTR (#61), Kimura 1980 (#80), and Jukes Cantor (#88) nucleotide substitution models. For amino acid topologies (MB_AA and TT_AA), sequences were inferred under the JTT+G (ProtTest rank #8), JTT (#41), and DAYHOFF (#86) amino acid substitution models.

3s.1 Resurrections:

All ML inferences were resurrected in the laboratory except for the sequences inferred under the JC and the GTR models.

## References


1. Posada, D. and K.A. Crandall, *MODELTEST: testing the model of DNA substitution.* Bioinformatics, 1998. **14**(9): p. 817-818.
2. Posada, D., *jModelTest: Phylogenetic Model Averaging.* Molecular Biology and Evolution, 2008. **25**(7): p. 1253-1256.
3. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21**(9): p. 2104-5.
4. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
5. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.

# CHAPTER 5

# THE FLUORESCENT PROTEIN EXPERIMENTAL PHYLOGENY


## Abstract

An experimentally derived fluorescent protein (FP) gene phylogeny starting from a single mRFP gene sequence was constructed in the laboratory. The mutated descendant sequences were evolved according to a scenario in which adaptive/diversifying evolution occurs along branches of the tree and with convergent evolution generating sub-lineages whose subsequent progeny fluoresce in the same spectrum but different from the ancestor. The branch lengths of the experimental phylogeny are evolving in a way that corresponds to branches associated with natural fluorescent proteins as permitted by the experimental system. Adaptive and purifying selection of the evolved descendants is determined by visually inspecting bacterial plates under the appropriate UV lighting conditions. The generated mRFP experimental phylogeny is represented in Figure 5.2. Figure 5.4 shows the phylogenetic tree drawn with bacteria expressing the phenotypes associated with each branch and node.


## Constructing a Functionally Diverse Protein Phylogeny

Generating an experimental phylogeny requires two essential components: 1) a sequence and 2) a mutagen to evolve the sequence. The

phylogeny in this study was generated from a single monomeric red fluorescent protein (mRFP 1.0) gene sequence that has evolved through mutations introduced by an error-prone DNA polymerase. Specifically, the mRFP 1.0 (parent) was amplified in a random mutagenesis PCR whereby most of the amplified gene products (descendents) contained a unique set of mutations, provided that the error-prone polymerase incorporated a non-native nucleotide at a particular site or sites. This descendant gene population was then optimized for cloning into an expression vector and transformation into an *E. coli* expression host (see Methods Section for details). Figure 5.1 is a representation of the descendant population's phenotype. Mutations that alter a FP's phenotype can be visualized when the protein is expressed within an *E. coli* colony under a UV light source (365 nm).

Next, descendants were then selected based on their phenotypes as determined under a UV lamp (365 nm), and then sequenced to determine synonymous and nonsynonymous mutations. One to two descendants were selected after each mutagenesis in the phylogeny building process, and these descendant sequences then act as the 'parental DNA' for a subsequent random mutagenesis PCR round. The entire phylogeny is built upon the mutation and selection of descendant FP sequences (Figure 5.2, Figure 5.3A-B)

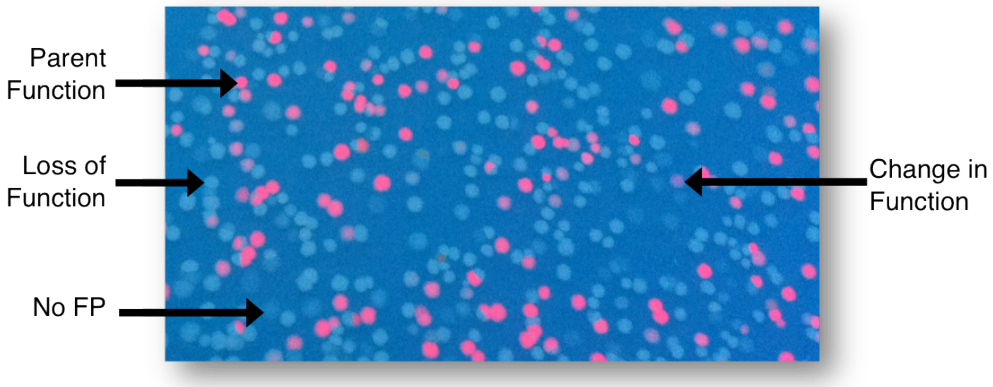



**Figure 5.1. Phenotypes of a mutant (descendent) population generated from mutagenesis of a mRFP red variant.** BL21(DE3) cells on a LB/CARB/IPTG/Agar plate expressing a mRFP mutant population whereby a mRFP red variant underwent one round of PCR random mutagenesis. Fours types of colonies are observed after each round mutagenesis as depicted by the arrows in the image: Colonies are representative of the parent phenotype, a different phenotype, a loss of phenotype, or colonies that do not express a FP gene at all. 'No FP' refers to a colony that is not expressing protein, while 'Loss of Function' refers to a colony expressing protein that has lost its function.

## Understanding the Evolved Phylogeny

The evolution of an experimental phylogeny is controlled by its source of mutation. The mutagen evolving the mRFP sequence into a functionally diverse phylogeny was an error-prone PCR polymerase, Mutazyme® II DNA polymerase. Mutazyme® II DNA polymerase is an ideal choice for evolving our mRFP sequences in a biologically consistent manner because it incorporates substitutions during DNA replication, a common mechanism among all biological organisms and the cause of most genetic mutations within organisms. We also have to consider the specific types of nucleotide substitutions the polymerase

favors. Ideally we want an unbiased polymerase to generate random mutations, or mutations of biological significance, meaning that the mutational bias of the polymerase should reflect the nucleotide substitution patterns of real biological organisms. In addition, we are only selecting for functional proteins, and this selection bias should contribute to a more biologically consistent mode of evolution. An overview of Mutazyme® II DNA polymerase's mutational spectra is provided in Table 5.1 and compared to the mutational properties of a natural mRFP phylogeny.

Stratagene provides us with a variety of ways way to assess the mutational bias exhibited by Mutazyme® II DNA polymerase. One bias indicator is determined by analyzing transition and transversion ratios (Ts/Tv). There are four possible transition and eight possible transversion mutations that can occur in DNA sequences. Therefore a truly unbiased enzyme should have a Ts/Tv ratio of 0.5. Mutazyme® II's Ts/Tv shows a bias towards transition mutations while the evolved mRFP sequences highly favor transitions over transversions with a ratio of 1.55. Within biological systems, transitions are actually favored over transversions since a transition do not alter the encoded protein sequence, thus Mutazyme® II's and the mRFP phylogeny's transitional bias is consistent with real organismal evolution. Setting aside our different Ts/Tv ratios, the experimental phylogeny is very similar in all transition and transversion values, thus the experimental phylogeny either is representative of the polymerase's mutational spectra, or Mutazyme® II's mutational spectra is truly representative

**Table 5.1. Mutational spectra of an entire mutant population versus selected variants of the experimental fluorescent protein phylogeny.** Stratagene provides Mutazyme® II DNA polymerase's mutational spectra, and this spectra is a representation of an entire mutagenesis reaction. The experimental mRFP phylogeny only represents a subset of the mutant population that includes only one to two sequences from each round of mutagenesis.

| Type(s) of mutations | Mutant | Selected Mutants |
|---|---|---|
| **Bias Indicators** | | |
| **Ts/Tv** | **0.9** | **1.55** |
| **AT → GC / GC → AT** | **0.6** | **0.70** |
| **A → N, T → N** | **50.70%** | **47.80%** |
| **G → N, C →** | **43.80%** | **52.20%** |
| **Transitions** | | |
| **A → G, T → C** | **17.50%** | **24.90%** |
| **G → A, C → T** | **25.50%** | **35.90%** |
| **Transversions** | | |
| **A → T, T → A** | **28.50%** | **20.30%** |
| **A → C, T → G** | **4.70%** | **2.70%** |
| **G → C, C → G** | **4.10%** | **4.90%** |
| **G → T, C → A** | **14.10%** | **11.40%** |
| **Mutation Frequency** | | |
| **Mutations/kb (per** | **3-16 (per PCR)** | **3-4 (per PCR)** |

of biological reality since the proteins are evolving biological sequences. Another bias indicator can be assessed by calculating the ratio of AT -> GC to GC-> AT transition mutations, which would be equal for an unbiased enzyme. At a value of 0.6, we see that Mutazyme® II favors GC transitions over AT transitions. Lastly, bias can be assessed by comparing the frequency of mutating A's and T's versus

the frequency of mutating G's and C's, which should be equal for an unbiased polymerase. We see that these values are somewhat similar, with Mutazyme® II favoring mutations in A (changing to any base but itself) and T (changing to any base other than itself). The error prone polymerase favors G and C transitions over A and T transitions and highly favors A → T / T → A transversions.

It is important to note that Mutazyme® II's mutational spectra is represented as the average mutational spectra of the entire amplified population. However, the mRFP phylogeny only represents the mutational spectra of the one or two selected mutants displaying the desired phenotype. Specifically, our FP mutational spectra data is gathered from the entire FP phylogeny encompassing all mutations that have occurred between a parent and it's immediate offspring. Our selected mutants are not representative of the entire mutagenesis reaction because our selection method is biased on selecting for a functional mutant FP. This explains why the average mutations per kb per PCR reaction of our FPs (3-4 mutations per PCR) are far fewer than that representing the entire population (3-16 mutations per PCR) because mutants containing an average of 16 mutations per kb result in a non-functional or 'dead' FP.

Overall, the mechanism by which Mutazyme® II DNA polymerase incorporates mutations is biologically consistent thereby allowing us to build and evolve an experimental phylogeny with more biological realism than those of computer simulations.

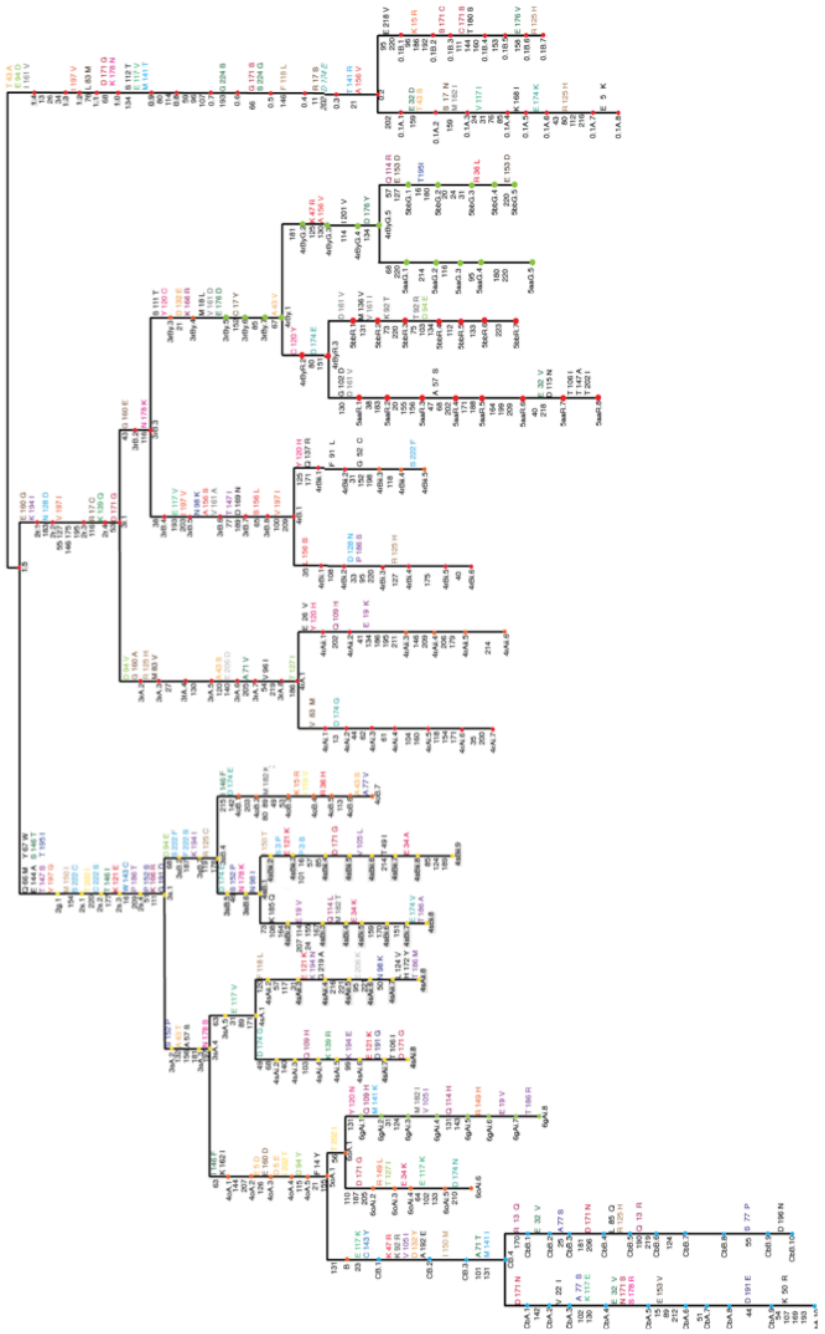


**Figure 5.2. A phylogeny of mRFP variants.** 1.0 (see Figure 53.B for clarity) represents recombinant mRFP used by the research community for microscopy imaging while every other circle represents a variant selected at the end of a random mutagenesis experiment. The colors of the circles indicate the phenotype (color emission) of the variant fluorescent protein (blue, red, green, yellow and orange). All mutations separating nodes are listed. Synonymous mutations are listed to the left of a branch, while nonsynonymous mutations are listed to the right of the branch. Multiple nonsynonymous mutations at a single site are highlighted in the same text color but have no correlation to phenotype.

Figure 5.2 is a representation of our experimental phylogeny. The root of the tree, node 1.5, is the last common ancestor (LCA) of the experimental FP phylogeny (Figure 5.2, node and variants can be seen in Figure 5.3A and 5.3B). This ancient node connects the in-group portion of the tree to the two out-group sequences 0.1A.8 and 0.1B.7, and has a red phenotype. Subsequent evolution of this red phenotype gave rise to the diverse spectra of colors found within the experimental FP phylogeny. (Figure 5.4 & Table 5.2). There are a total of nineteen taxa that represent the terminal branches of the phylogenetic tree, with blue (CbA.10 & CbB.10), green (6oAi.8, 5aaG.5, & 5bbG.5), yellow (4sAi.8, 4sAii.8, 4sBi.8, & 4sBii.9), orange (6oAi.6, 4oB.7, 4rAii.6, & 4rBii.5), and red (4rAi.7, 4rBi.6, 5aaR.8, 5bbR.7, 0.1A.8, & 0.1B.7) phenotypes.

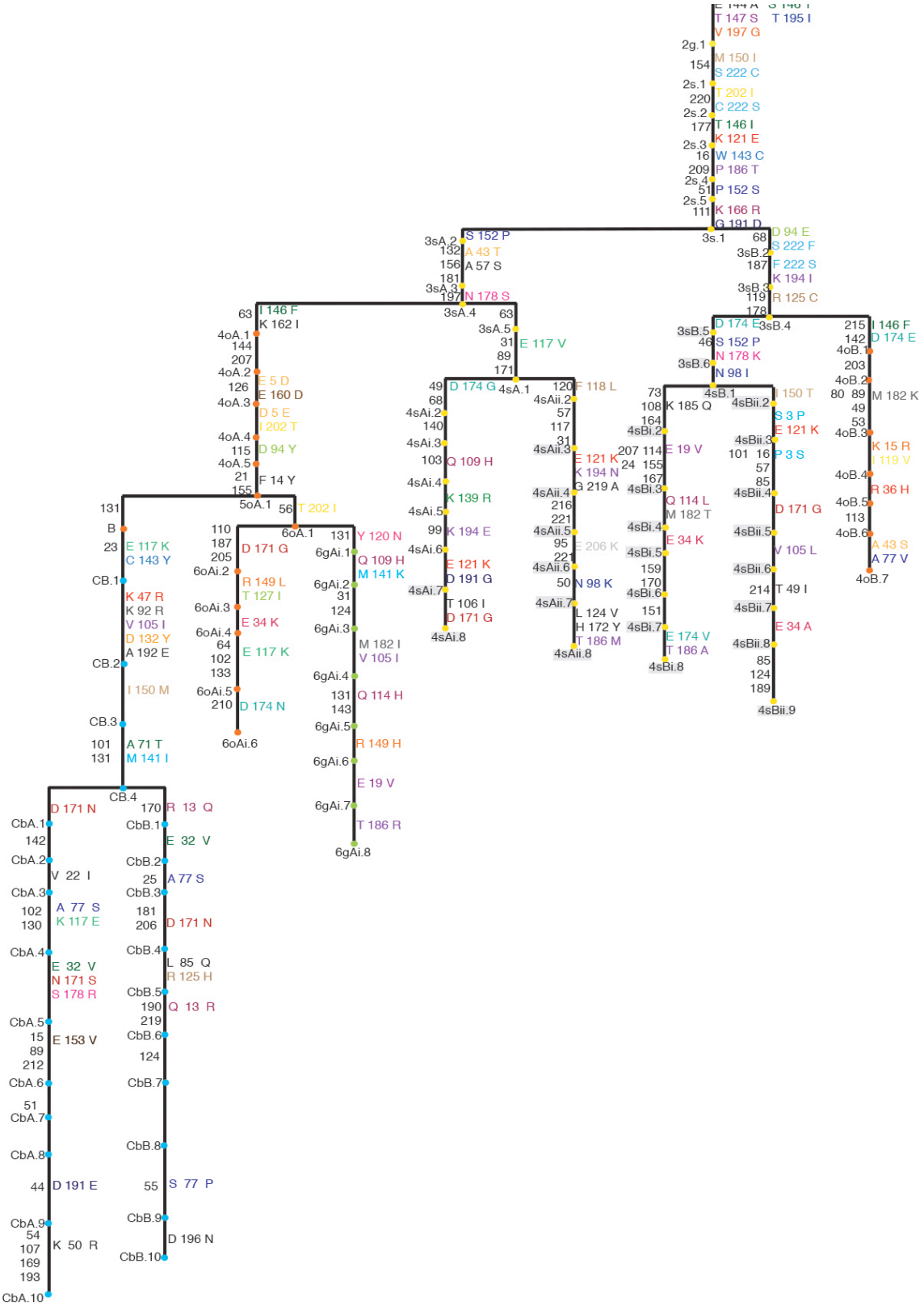


**Figure 5.3A. Close-up left half portion of the tree depicted in Figure 5.2.** This portion represents the left half of the tree, while the right half of the tree is depicted In Figure 5.3B.

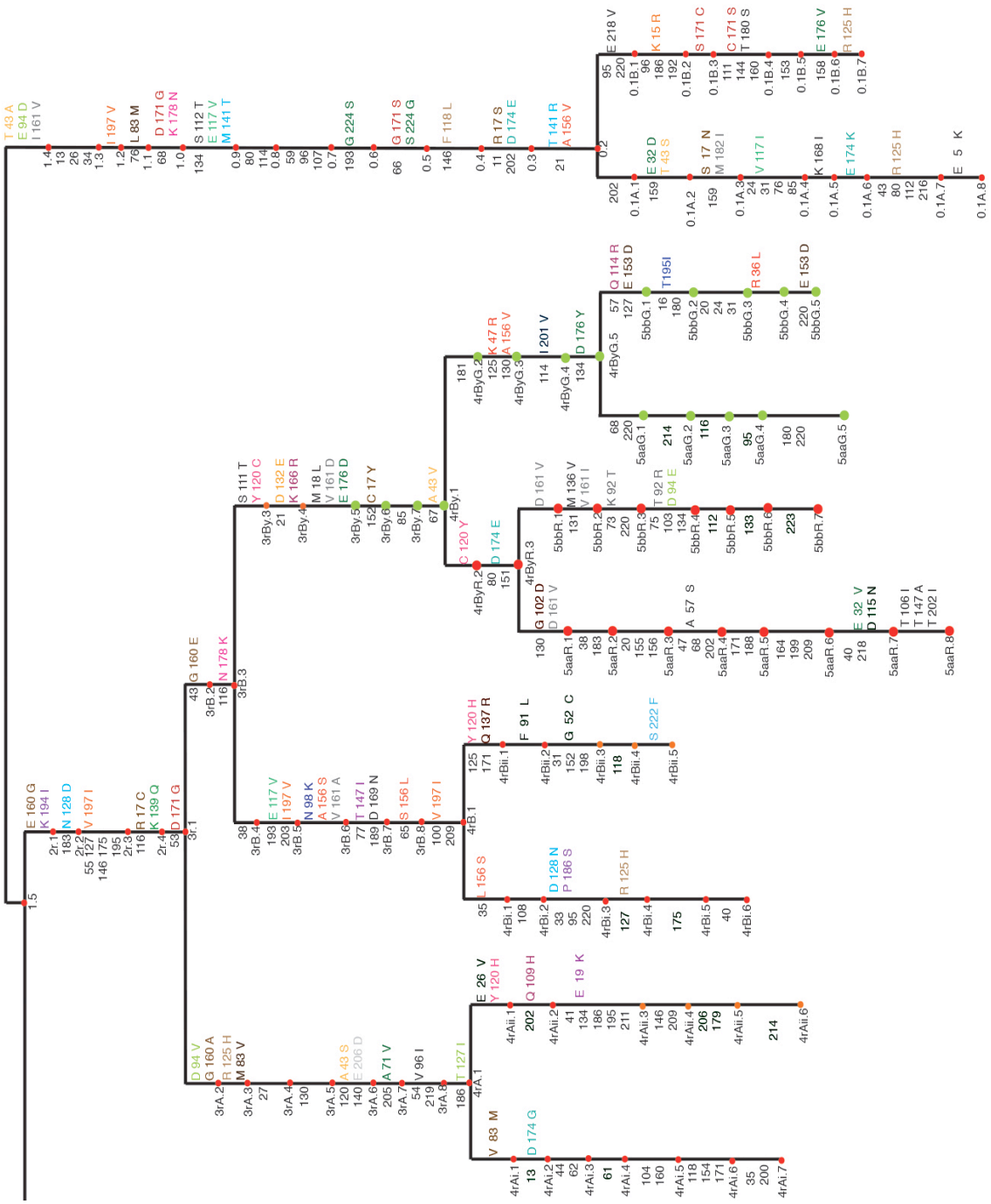


**Figure 5.3B. Close-up right half portion of the tree depicted in Figure 5.2.** This is the right half of the tree. The left half is depicted in Figure 5.3A.

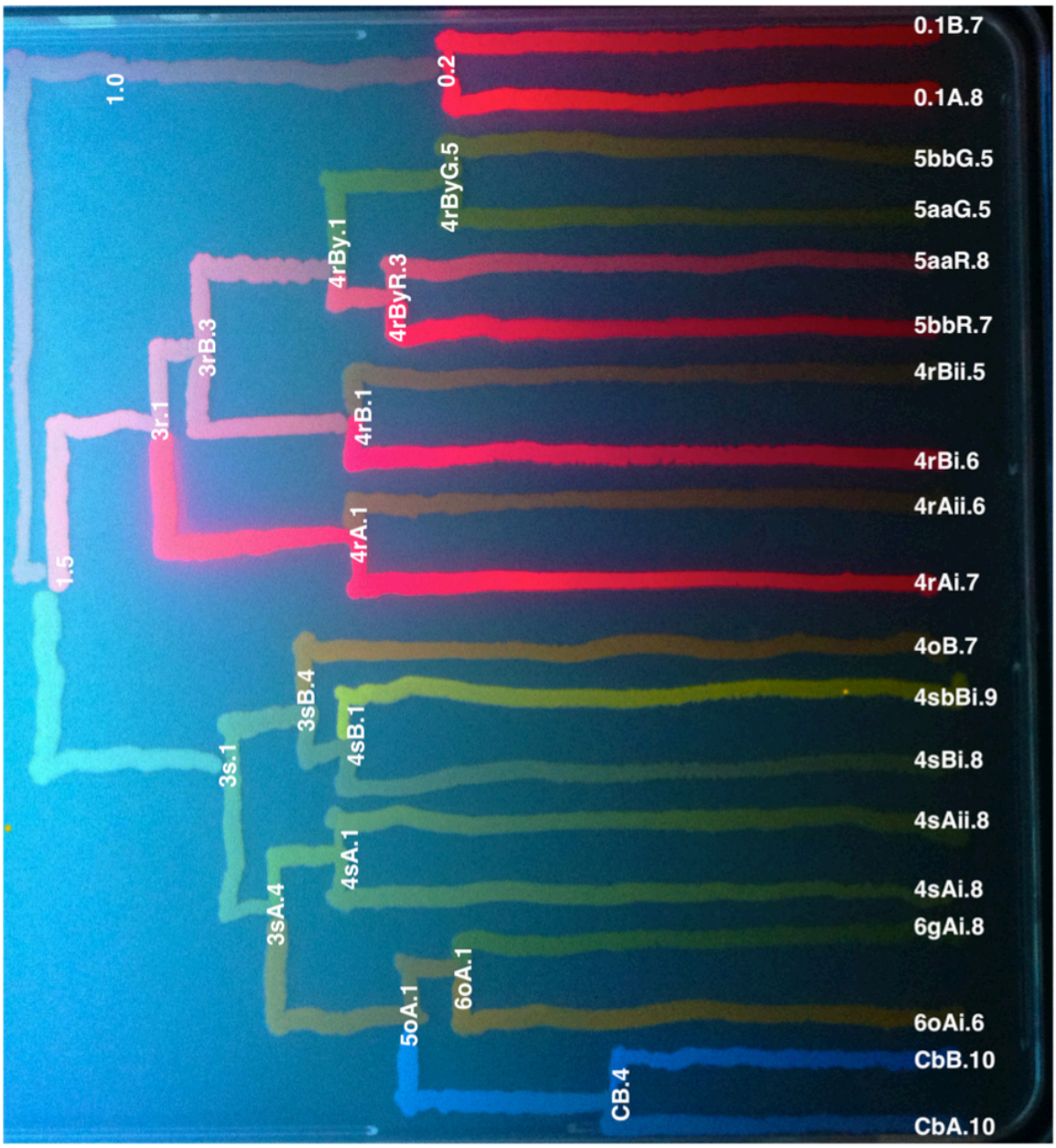



**Figure 5.4. Experimentally derived mRFP Cladogram.** Cladogram drawn with bacteria growing on an agar plate and expressing actual fluorescent protein variants evolved during our experimental phylogeny and visualized under ultraviolet light. Topology is a representation of true topology shown in Figure 5.2. The experiments started with a single monomeric red fluorescent protein (labeled as 1.0) and from this red phenotype we have evolved green, yellow and orange color emission phenotypes. White labels at the leaves of the tree denote the name of the "extant" proteins. White labels at the nodes of the tree denote the name of the ancestor at that node. Branch colors are the actual phenotypes of the labeled nodes and terminal sequences. This figure conveys the diversity of phenotypes evolved in our experimental phylogeny. Internal branches depict ancestral phenotypes of blue, green, yellow, orange, and red.

## Characteristics and Episodes of Functional Divergence

Our experimental phylogeny has evolved in such a manner that some portions of the tree are experiencing diversifying selection while other portions are experiencing purifying selection (Figure 5.2). Sometimes, the diversifying selection was the result of a burst of amino acid replacements against a background of otherwise few synonymous substitutions (high dN/dS, Ka/Ks ratio) (Figure 5.3A [1.5 → 3s.1]), while other portions of the tree contain diversifying selection within a background of equal nonsynonymous and synonymous substitutions (Figure 5.3A [3sB.4 → 4oB.7]). We also propose to evolve the tree to contain examples of convergent evolution (both at the phenotypic level and site-specific level discussed later) since this is biologically relevant and known to confound most phylogenetic algorithms. How such convergence is handled by ASR approaches is unknown.

In general, FPs contain a broad range of genetic variants whose emission profiles span nearly the entire visible spectrum. FPs are generally divided into seven spectral classes based upon their emission maxima. This includes FPs emitting in the blue (BFPs; 440–470 nm), cyan (CFPs; 471–500 nm), green (GFPs; 501–520 nm), yellow (YFPs; 521–550 nm), orange (OFPs; 551–575 nm), red (RFPs: 576–610 nm) and far-red (FRFPs; 611–660 nm) spectral regions. In addition to the conjugated ring system itself, the local protein micro-environment surrounding the chromophore can also have an impact on the spectra [1].

Table 5.2 summarizes the spectral characteristics of the evolved proteins in this study and shows all of their excitation and emission spectra.

**Table 5.2. Properties of terminal and ancestral fluorescent proteins***. Images are pictures of bacteria expressing FPs.

| Phenotype | True Name | Excitation max (nm) | Emission max (nm) | Chromophore |
|---|---|---|---|---|
|  | CbA.10 | 453 | Peak on or before 474 | MWG |
|  | CbB.10 | 453 | Peak on or before 474 | MWG |
|  | 5aaG.5 | 488 | Peak on or before 496 | QYG |
|  | 5bbG.5 | 488/514 | Peak on or before 496 (minor 528) | QYG |
|  | 6gAi.8 | 488/514 | Peak on or before 506 (minor 528, 570, 581) | MWG |
|  | 4rBy.1 | 488/514 | Peak on or before 496 (minor 528) | QYG |
|  | 4sAi.8 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 4sAii.8 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 4sBi.8 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 4sBii.9 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 3s.1 | 488/514 | 528 | MWG |
|  | 4oB.7 | 488/514 | 517 and 570 | MWG |
|  | 4rAii.6 | 488/514 | Peak at or before 506 (minor 549) | QYG |
|  | 6oAi.6 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 4rBii.5 | 488/514 | Peak on or before 506 (minor 549) | QYG |
|  | 5oA.1 | 488/514 | 517, 570 and 592 (minor 645) | MWG |
|  | 4rAi.7 | 488/514 | 592 | QYG |
|  | 4rBi.6 | 543 | 581 | QYG |
|  | 5bbR.7 | 543 | 581 and 592 nm | QYG |
|  | 5aaR.8 | 543 | 581 nm | QYG |
|  | 0.1A.8 | 543 | Broad peak range 581-592 | QYG |
|  | 0.1B.7 | 543 | Broad peak range 581-592 | QYG |
|  | 1.5 | 543 | Single Peak at 592 | QYG |
|  | 3r.1 | 543 | Broad peak range 581-592 | QYG |

*selected ancestors representing nodes of interest based on functional divergence. The chromophore column shows the three amino acid residues that make up the precursor of the chromophore.

The following is a description of the general characteristics of the experimentally evolved spectral classes of FPs:

**Blue**

The experimentally derived FP phylogeny contains two terminal leaves each representing a novel blue FP, CbA.10 and CbB.10 (Figure 5.4 and Figure 5.5). The most recent common ancestor of CbA10 and CbB10 is CB.4, which also has a blue phenotype (Figure 5.3A). CB.4 is a descendent of the orange 5oA.1 ancestral node. As such, the functional change leading to the blue phenotype occurred along the branch that leads from the orange 5oA.1 node to the younger blue CB.4 node. Specifically, two sequential rounds of evolution leading to the functional change were observed. Orange 5oA.1 was mutagenized and its direct descendant, orange variant B (Figure 5.3A), was selected. Orange variant B is different from its parent 5oA.1 by one synonymous mutation at residue 131. Orange variant B was then put through one mutagenesis cycle and its direct descendant, blue variant CB.1, was selected. CB.1 is different from its orange B parent by one synonymous change at residue 23 and two nonsynonymous changes at E117K and C143Y. CB.1 underwent further evolution through three rounds of mutagenesis while maintaining its blue fluorescence (as depicted by extension of the branch from CB.1 to CB.4 by 2 synonymous and 8 nonsynonymous mutations in Figure 5.3A).

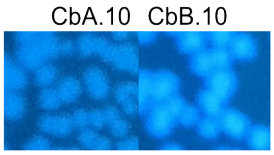



**Figure 5.5. Phenotypes of CbA.10 and CbB.10.** Bacterial colonies on an LB/CARB/IPTG/Agar plate expressing CbA.10 (image on left) and CbB.10 (image on right) FPs.

CB.4 then underwent a 'speciation' event, as designated by the bifurcation event at node CB.4, resulting in two blue fluorescent sub-lineages. The two blue sub-lineages evolved having equal synonymous to nonsynonymous substitutions (total mutations along the A branch is 24 [12 synonymous and 12 nonsynonymous substitutions] and 17 along the B branch [8 synonymous and 9 nonsynonymous substitution]). Although this pattern may be indicative of neutral-like evolutionary processes, purifying selection was maintained in the sense that the blue phenotype itself was maintained. Again, the 5oA.1 orange ancestor switched to a blue emitting protein through the accumulation two nonsynonymous mutations: E117K and C143Y. The appearance of blue is an ideal example conveying the connection of genotype to phenotype wherein a genotypic change directly results in a functional change. The more bulky and aromatic tyrosine replaced a cysteine at site 143. Site 143 is known to directly interact with the

helix that contains the chromophore triad, thus influences the local environment of the chromophore and affects color emission.

It is intriguing to speculate that site 143 is responsible for the conversion from the orange phenotype to the blue phenotype by simply replacing the cysteine residue with a bulky aromatic tyrosine.(Figure 5.6 and 5.7). Curiously, however, the mRFP protein also has a bulky aromatic residue at this position, tryptophan. We speculate that the conversion of tryptophan to tyrosine may change red phenotype to blue, or that the tyrosine in blue has epistatic interactions with other amino acid residues present in orange that are not present in red and that these interactions give rise to the blue phenotype.

The blue variants also serve to highlight the population of descendants generated from a PCR mutagenesis reaction whereby the descendants are ligated into pET-15b vector and then transformed into E. coli (Figure 5.8). The figure shows that the vast majority of colonies have a fluorescent protein such that most of these colonies are blue but a substantial portion are also variants of blue - in this case yellowish. Very few of the colonies represent pET-15b vector that closed in on itself without ligating a fluorescent gene

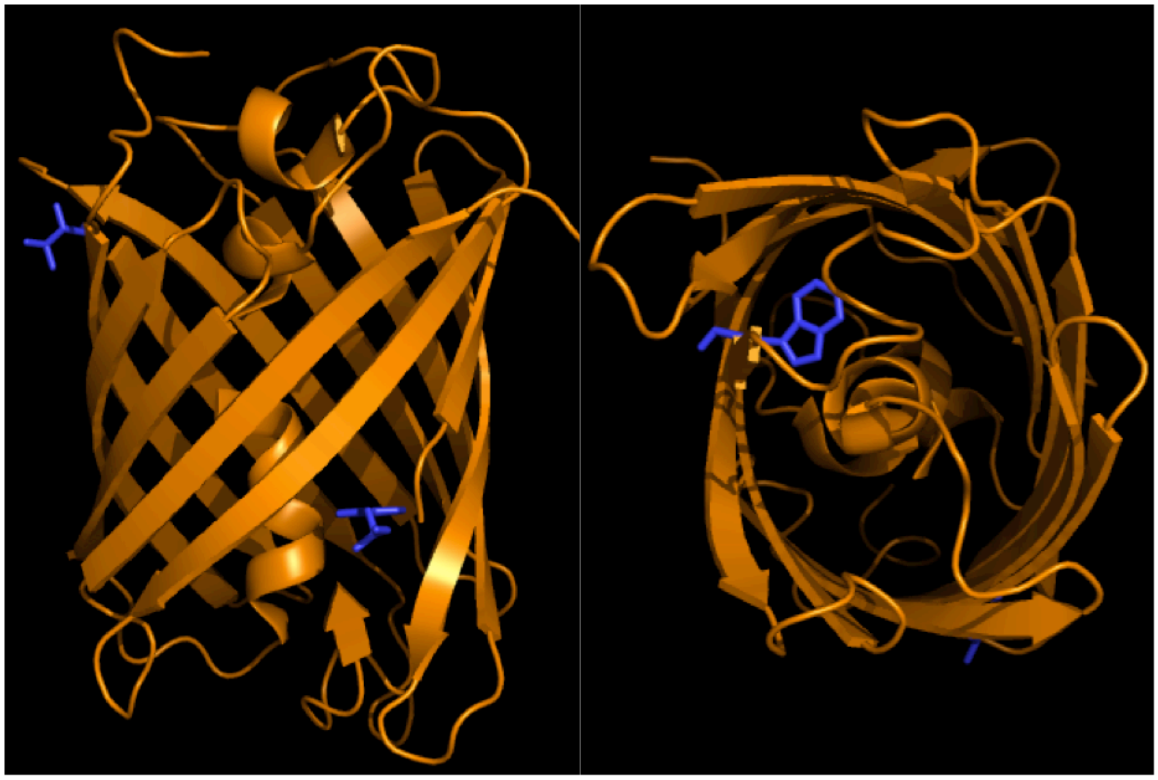


**Figure 5.6. Conversion of orange 5oA.1 to blue CB.1.** This image depicts the structure of mRFP1.0 (PDB ID 2VAD) with amino acid residues 143 (Tryptophan, W) and 117 (Threonine, T) from mRFP1.0 highlighted in blue. The figure on the right represents the transverse view of mRFP1.0 indicating the possible chemical interactions of the aromatic tryptophan residue with the chromophore. In the mutation of blue from orange, the cysteine interactions at site 143 are replaced by the interactions of the aromatic tyrosine and this is hypothesized to lead to the phenotypic change.


**Figure 5.7. Multiple sequence alignment of mRFP 1.0, 5oA.1, and CB.1**. The starting-protein mRFP 1.0 and the ancestral proteins 5oA1 and CB.1. Residues not identical between the three sequences are highlighted in orange.

Curiously, blue is the only FP within our experimental phylogeny that has not randomly evolved more than once. Further, blue variants are the least-robust phenotypes within the phylogeny, indicating that random mutagenesis of blue variants nearly always generates non-blue phenotypes in the descendent population whereas all other FPs are more difficult to change phenotypes. We therefore conclude that the blue phenotype possibly exists in a narrow sequence

space and that it is easy to leave this sequence space through many different mutational pathways.

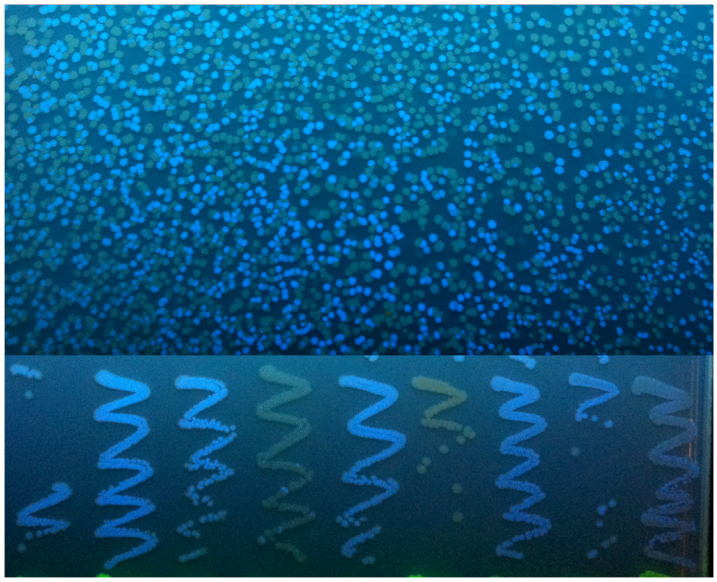


**Figure 5.8. mRFP transformation and selection of mutagenized mRFP blue variant.** Top portion of the figure represents the plating onto LB/carb/IPTG of a bacterial transformation of a PCR mutagenesis of a blue fluorescent gene ligated into the pET-15b expression vector. The bottom portion of the figure shows streaks of single colonies picked from the top portion of the figure. The steaks that are not blue (but yellowish) are mutant phenotypes.

The spectral profiles of CbA.10 and CbB.10 were determined using equipment at Georgia Tech (Figure 5.9).

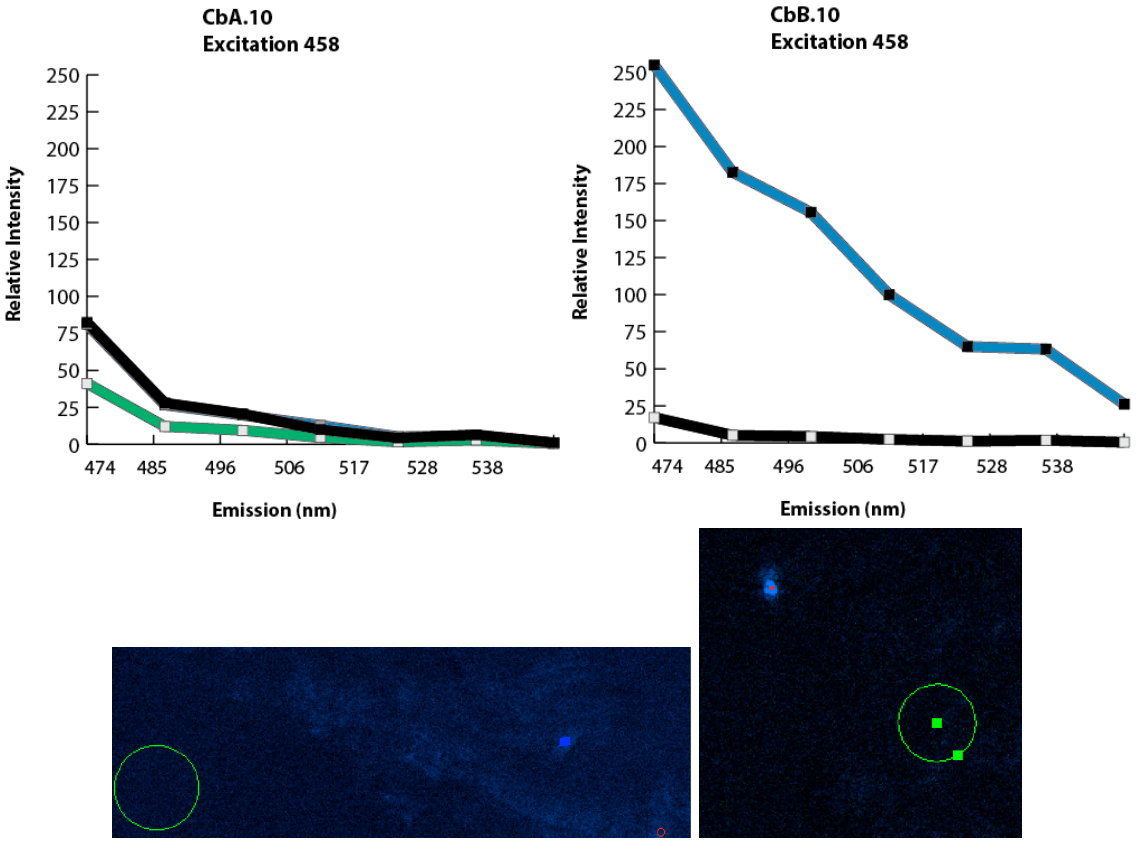



**Figure 5.9. CbA.10 and CbB.10 spectra and microscope images.** Top left -Black line is the emission spectra of CbA.10 excited at 458 and the green line represents background. Top right - Blue line is the emission spectra of CbB.10 excited at 458 and the black line represents background. Bottom left and right - Green circles represent the area of background selected for emission readings of CbA.10 & CbB.10. Blue dot in the image represents bacteria selected for spectral reading. Image on the left had no clear bacteria in focus and hence the best available area was selected whose spectra is depicted in CbA.10. Blue dot on right image is a bright bacteria in focus and this bacteria's spectra is depicted in spectra CbB.10. Images were taken on a NLO Meta Microscope and captured using Zen 2008 software. This equipment only allowed for the measurement of the emission wavelengths at 474 nm or greater with the lowest wavelength of excitation light at 458 nm. As can be derived from the spectral curves, the emission peaks are on or before 474 nm but the specific values cannot be determined. CbA.10 has low fluorescence and high background.

## Green

There are two recorded instances where green phenotype randomly evolved within the FP experimental phylogeny. The first instance that green

phenotype occurred is at the bifurcation of the orange ancestral node 6oA.1(Figure 5.3A). This phenotypic change occurred due to a single nonsynonymous substitution at site 120 (tyrosine replaced asparagine) and a synonymous substitution at residue 131. Further, site 120 is shown to directly interact with the FP's internal environment as depicted in Figure 5.10, which leads us to believe that the change in the phenotype from orange to green is the direct result of the T120N replacement.

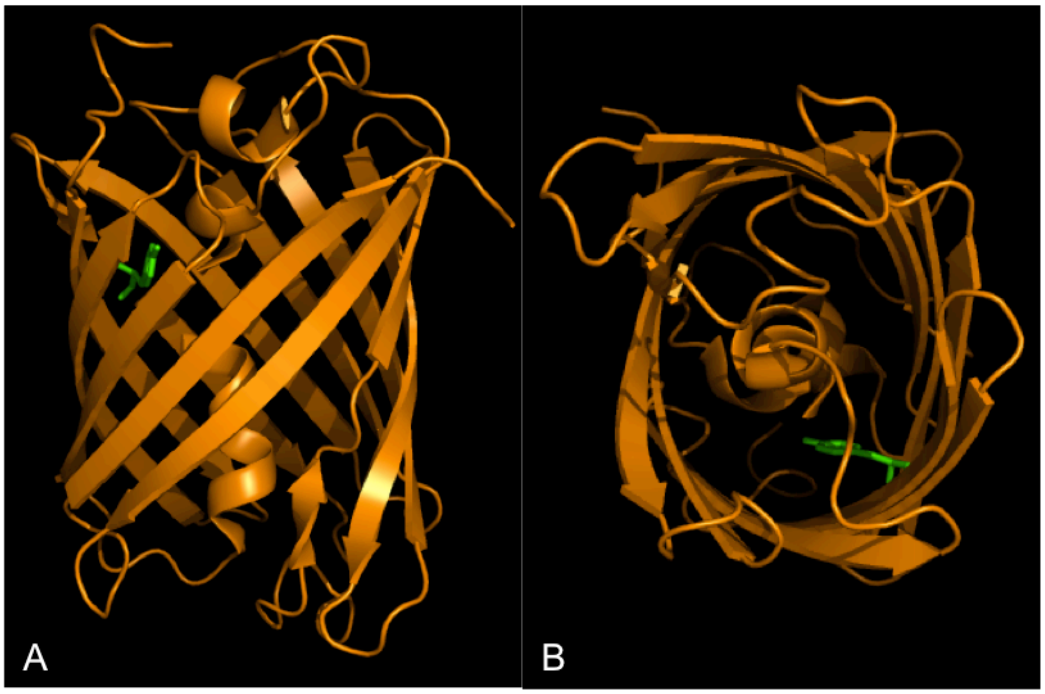



**Figure 5.10. Conversion of orange 6oA.1 to green 3rBy.5.** This image depicts the structure of mRFP 1.0 (PDB ID 2VAD) with amino acid residue 120 (Tyrosine, Y) from mRFP1.0 highlighted in green. A**.** Side view of mRFP 1.0. B. Traverse view of mRFP1.0 indicating the possible chemical interactions of the aromatic tyrosine residue with the chromophore environment. In the mutation from orange to green phenotype, the tyrosine interactions at site 120 are replaced by the interactions of asparagine, or lack there of, and this is hypothesized to lead to the green shifted phenotypic change.

The second instance of green phenotype is seen along the branch connecting red ancestral node 3rB.3 to green ancestral node 4rBy.1 (Figure 5.3B). Red ancestor 3rB.3 endured one round of random mutagenesis. A direct descendant displaying a unique phenotype was then selected for sequential evolution and named 3rBy.3. It displayed a dual-like phenotype because it would display a yellow phenotype in some instances, and a reddish phenotype in other instances. The determinants for the color change were not assessed, however, the genetically encoded amino acid replacements S111T and Y120C were the primary cause of the 3rBy.3's phenotypic behavior. 3rBy.3 next endured one round of random mutagenesis, and its chosen offspring contained three nucleotide substitutions, one causing a synonymous change at residue 21, and the other two were responsible for D132E and K166R nonsynonymous substitutions; this sequence was named 3rBy.4 and its phenotype characteristics were similar to parent 3rBy.3. The mutagenesis of 3rBy.4 led to the first instance of green phenotype within the FP phylogeny: the emergence of green FP 3rBy.5 with amino acid replacements M18L, V161D, and E176D. Thus, 3rBy.5 green phenotype evolved from a red phenotype (3rB.3) within three rounds of mutagenesis, and it possessed an intermediate phenotype during its evolution.

3rBy.5 contains a total of seven nonsynonymous substitutions when compared with its immediate red ancestor 3rB.3. (Figure 5.11), and the specific locations of these residues within the FP are shown in Figure 5.12.

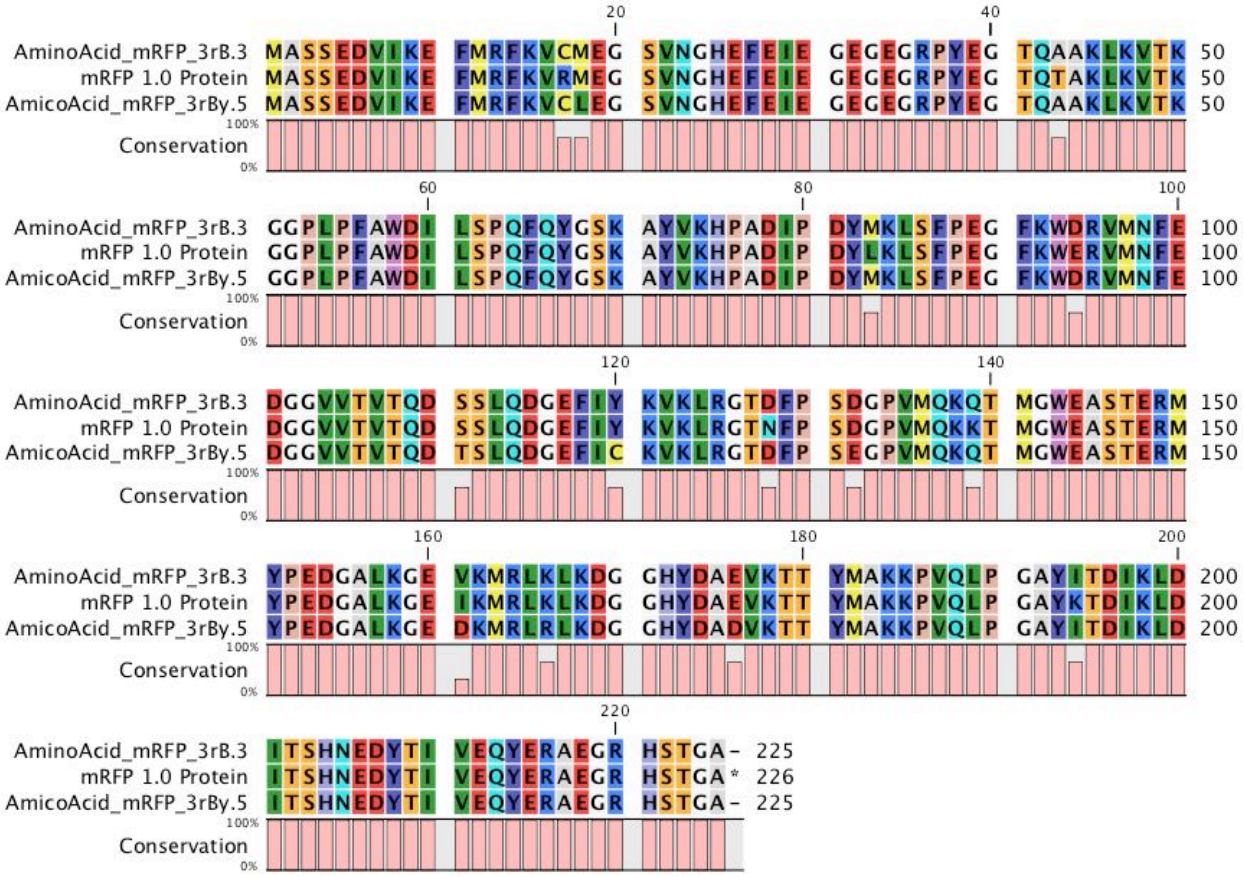



**Figure 5.11. Multiple sequence alignment of mRFP 1,0, 3rB.3, and 3rBy.5**. Short bars represent instances of mismatches.

To further understand the change in phenotype, we looked at mRFP's structure and highlighted the seven mutant residues in order to see if there were any obvious structural changes leading to the phenotypic change. We attempted to find the molecular determinants responsible for the color change by looking at the structure of the protein (Figure 5.12). The locations of the mutated residues within the FP protein were analyzed. The residue sites 120 and 18 seemed to play important roles in the interaction with the chromophore as they point towards

the interior of the protein. We also looked at the phylogeny's evolutionary history to determine if sites 111, 120, 161, 166, 176, 18 and/or 132 had been mutated before, and if they had, were the sites involved in an instance of color change. It has been observed within the experimental phylogeny that mutations at site 120 tend to cause a phenotypic color change. This can be explained based on the fact that site 120 is close to the chromophore region and thus, its bonding interactions affect the interior protein environment. In mRFP 1.0 and the 3rB.4 ancestor, site 120 is occupied by the aromatic amino acid tyrosine and site 18 consists of methionine. The first instance of color change occurs from 3rB.4 to the intermediate yellow/red phenotype because of the tyrosine to cysteine replacement at site 120 (Y120C).

In this study of experimental phylogeny, certain trends have been observed in regards to the amino acid residues present at particular sites like 120. When tyrosine is present at site 120, spectral properties show a signature of red-shifted fluorescence. While, when tyrosine is substituted by non-aromatic and single chain amino acids like asparagine or cysteine, the spectral properties tend to be blue-shifted. This observation can be attributed to the change in the chemical interactions in the protein structure and internal protein environment of the chromophore, as tyrosine is a bulky group whose aromatic ring provides for different bonding interactions than the single chain amino acids.

All nonsynonymous substitutions that have occurred at site 120 within our FP phylogeny have been involved in a color change. Even more interesting is

that site 120 does not lead to a specific color but actually will either substantially blue shift or red shift FP spectra. This residue is an ideal example of instances of convergent evolution within the experimental phylogeny. The native mRFP 1.0 ancestor has a tyrosine present at site 120. Substitutions at this site from tyrosine to cysteine and histidine have led to phenotypic change from red to orange and substitutions from tyrosine to asparagine have resulted in a green phenotype from an orange ancestor. Acquisition of this blue shift is suggestive of convergent evolutionary mechanisms since similar biological traits are derived within independent lineages.

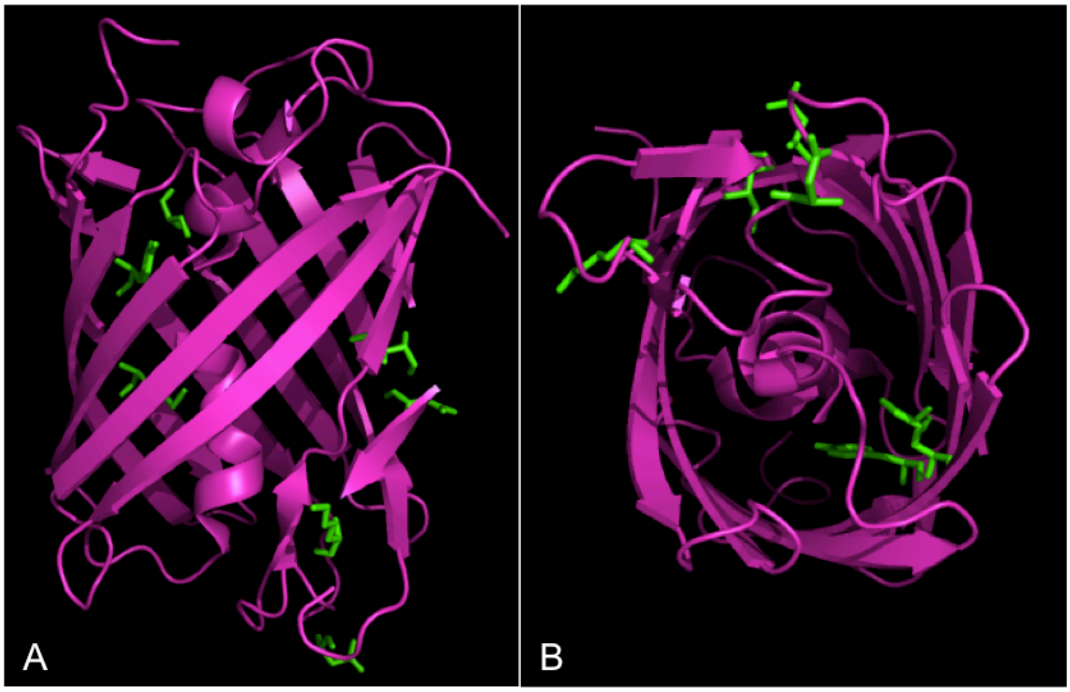



**Figure 5.12. A. B. Conversion of red 3rB.4 to green 3rBy.5.** This image depicts the structure of mRFP 1.0 (PDB ID 2VAD) with amino acid residues 111 (Serine, S), 120 (Tyrosine, Y), 132 (Aspartate, D), 166 (Lysine, K), 18 (Methionine, M), 161 (Isoleucine, I), and 176 (Aspartate, D) from mRFP1.0 highlighted in green. B. The transverse view of mRFP1.0 indicating the possible chemical interactions of the aromatic tyrosine residue with the chromophore.

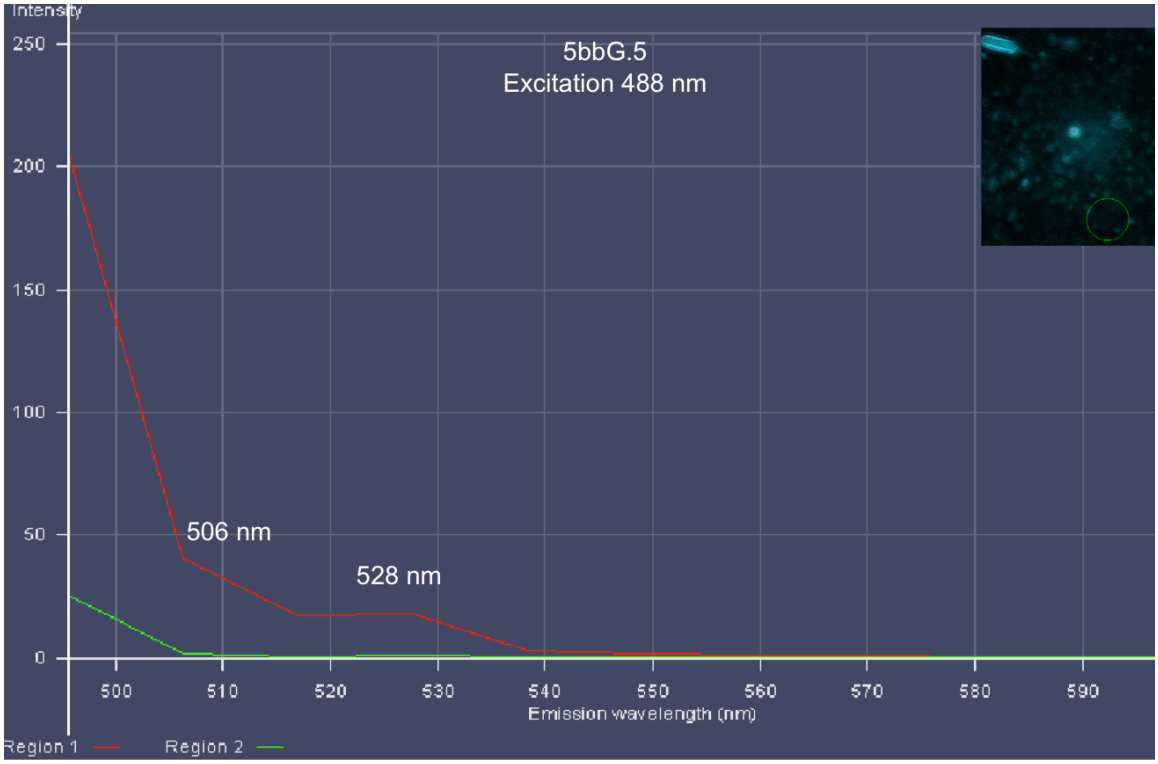



**Figure 5.13. Emission of green 5bbG5.** Excitation of 5bbG.5 at 488 nm leads to an emission maximum either at or before 495 nm. This finding suggests that 5bbG.5 is actually a Cyan fluorescent protein. The red curve gives the emission spectra of bacteria recombinantly expressing 5bbG.5 as depicted in the image in the upper left hand corner. The green curve is the spectra given off by the background signal. The green circle in the upper left hand image is the section selected for reading the background signal.

**Orange**

Ancestral 5oA.1 and its modern descendant 6oAi.6 are both orange FPs that are direct descendants of the yellow ancestral node 3sA.4 (Figure 5.3A). Thus the functional change leading from the ancestral yellow phenotype to the more derived orange phenotype occurred along the branch leading from the yellow 3sA.4 node to its orange 5oA.1 descendant. Specifically, orange occurred at the immediate bifurcation of node 3sA.4 into two lineages where one lineage possessed the same ancestral phenotype (yellow) and the other lineage possessed a new phenotype (orange). Orange variant 4oA.1 was selected from 3sA.4's mutant descendant population. Orange variant 4oA.1 is different from its parent 3sA.4 by one synonymous mutation at residue 63 and two nonsynomous mutations at residues 146 and 162 (Figure 5.14). Another instance when yellow evolved into the orange phenotype occurred when yellow ancestor 3sB.4 accumulated 2 synonymous mutations at residues 215 and 142, and 2 nonsynonymous mutations at residues 146 and 174 (Figure 5.15).

Instances of orange phenotype provide model examples of parallel evolution. The last common ancestor of 4oA.1 and 4oB.1 is 3s.1 (Figure 5.3A). 3s.1 has an isoleucine at position 146. 3s.1 gave rise to 4oA.1's most recent ancestor 3sA.4, and to 4oB.1's most recent ancestor 3sB.4. These ancestors still contain an isoleucine at position 146. Random mutagenesis of both ancestors resulted in a mutation at position 146 from an isoleucine to phenylalanine. This mutation caused the yellow ancestral phenotype to become orange (Figure 5.16).

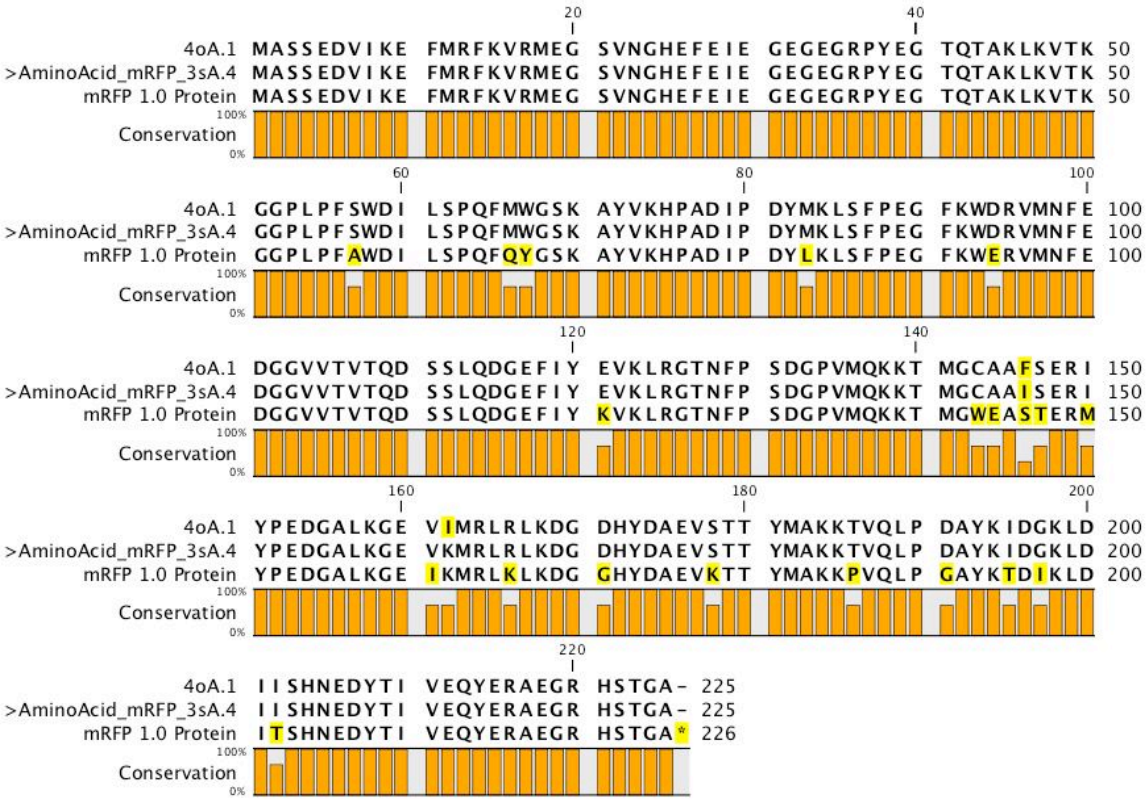



**Figure 5.14. Multiple sequence alignment of mRFP 1.0, 4oA.1, and 3sA.4**. Height of orange bars shows similarity i.e. tallest bars – all sequences have same residue at a site – shorter bars represent mismatches between the sequences. Yellow highlighted residues are sites that are different. Orange phenotype arose when ancestor 3sA.4 accrued 3 mutations in the 4oA.1 immediate descendant.

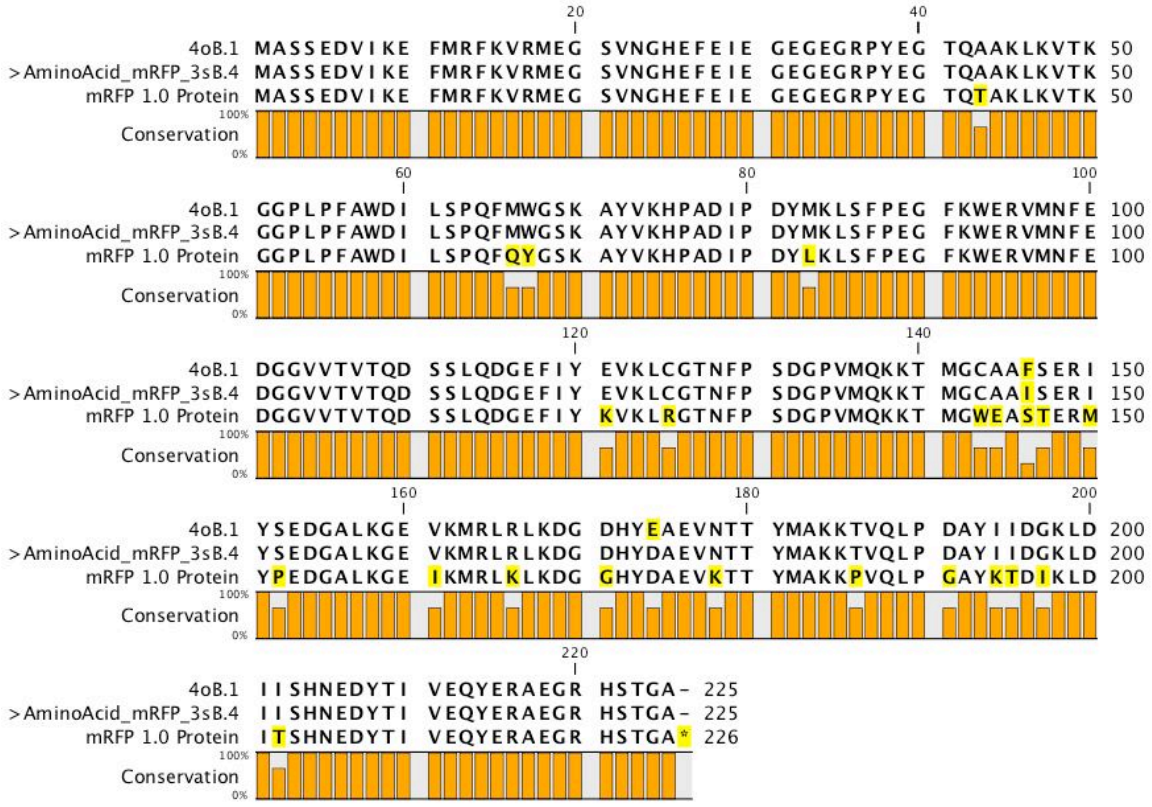



**Figure 5.15 Multiple sequence alignment of mRFP 1.0, 4oB.1, and 3sB.4.** Yellow phenotype in 3sB.4 changed to orange in 4oB.1 through the accumulation of two nonsynonymous mutations. 3sB.4 and 4oB.1 are compared with mRFP 1.0.

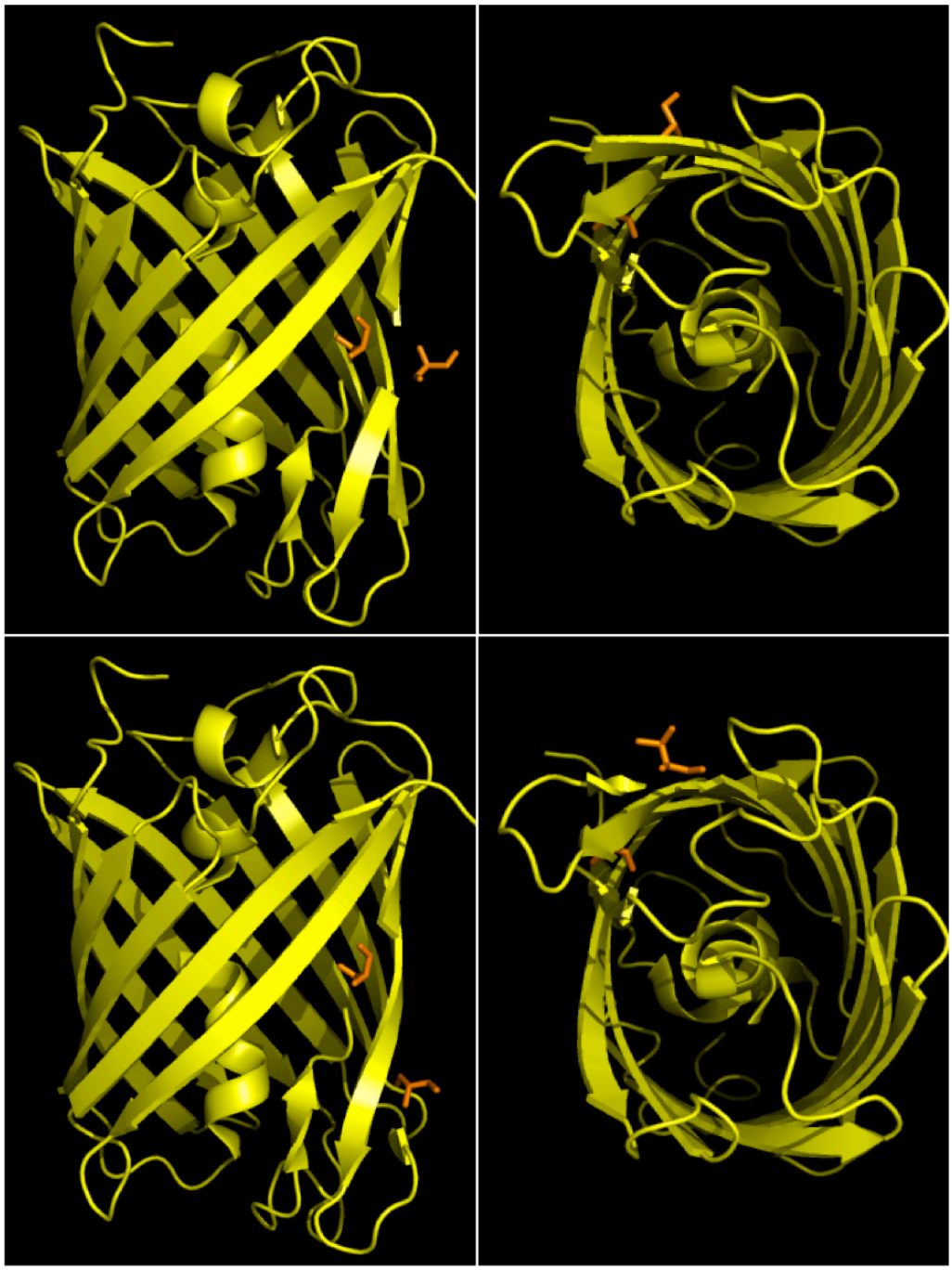


**Figure 5.16. 3sA.4 to 4oA.1 and 3sB.4 to 4oB.1.** Top: The PyMOL illustrations above represent the non-synonymous mutations of yellow FP in 3sA.4 to orange FP in 4oA.1. This phenotypic change is observed due to a mutation at site 146 from isoleucine to phenylalanine (I146F) and from lysine to isoleucine at site 162 (K162I). Bottom: This image shows the non-synonymous mutations of yellow FP in 3sB.4 to orange FP in 4oA.1. This phenotypic change is observed due to what we think is a singular mutation change at site 146 from isoleucine to phenylalanine (I146F), but also site 174 was mutated as well from aspartic acid to glutamate (D172E).

Microscopic spectral data reveal spectral diversity within the 6oAi.6 population (Figure 5.17). Orange bacteria are more red shifted that yellow bacteria. The green color in the image is due to high intensity which was a result of bacteria having high absorbance and being in focus. Red spectral curve represents area selected in an ‘orange’ bacteria. Blue curve is on ‘yellow’ bacteria. Yellow curve represents a spot with both ‘orange and yellow’ colors and the green curve represents an area of a ‘yellow’ bacteria. Multiple readings were taken because of the complexity of the spectral curves, and so as to fully characterize 6oAi.6’s spectra. Color variation may be due to multiple influences; locating of individual bacteria within colony, maturation of the FPs in different bacterial cells, etc.

When excited at 488 nm, 5oA.1 exhibits the same spectral emission curve as its descendant 6oAi.6. However, when 50A.1 is excited by the 514 nm laser, a new peak at 549 nm shows up. Spectra are consistent on all replicate readings. All bacteria show the same color under the scope, all look yellow, however upon close inspection each bacteria has red and green hues and look multicolored.

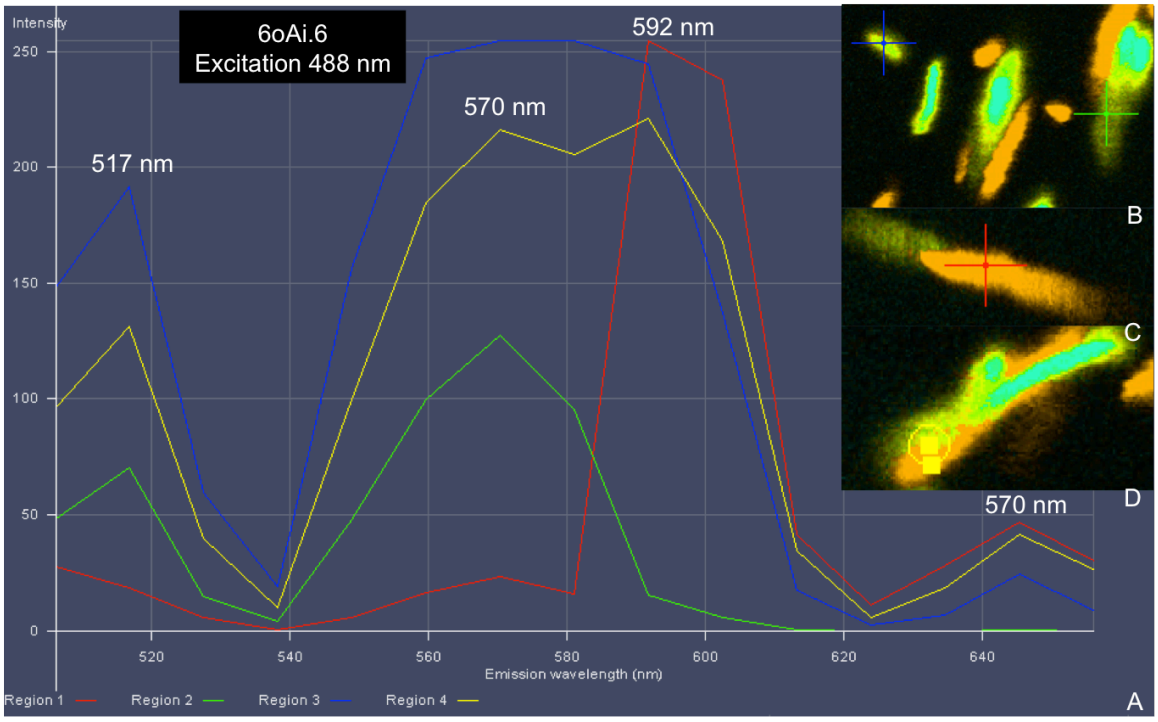



**Figure 5.17. 6oAi.6 spectra and microscope image.** A. Curves represent four spectral readings of 6oAi.6 excited at 488 nm taken all at once on the same sample (i.e. replicates). Peaks labeled. Green curve: peak at 517 and 570 nm. Red curve: 592 and 645 nm. Yellow Curve: 517, 570, and 592 and 645 nm. Blue curve: 517, 570 and 645 nm. B-D. Image of bacteria under microscope. Fluorescence was measured at the intersect point of the red, green, and blue crosses, and within the yellow circle area (image represents bacteria or section of bacteria selected for spectral reading). The curve color in A corresponds to the color of the cross or circle.. Bacteria are bright – high absorbance and in focus. Emission spectra before 506 nm could not be collected because lower nm would cause a high background signal and the 488 nm excitation laser would overlap – false signal. This image has no background signal. Note: COLOR OF BACTERIA DO NOT REPRESENT ACTUAL PHENOTYPE COLOR.

**5oA.1 Spectra: excitation 488 nm**

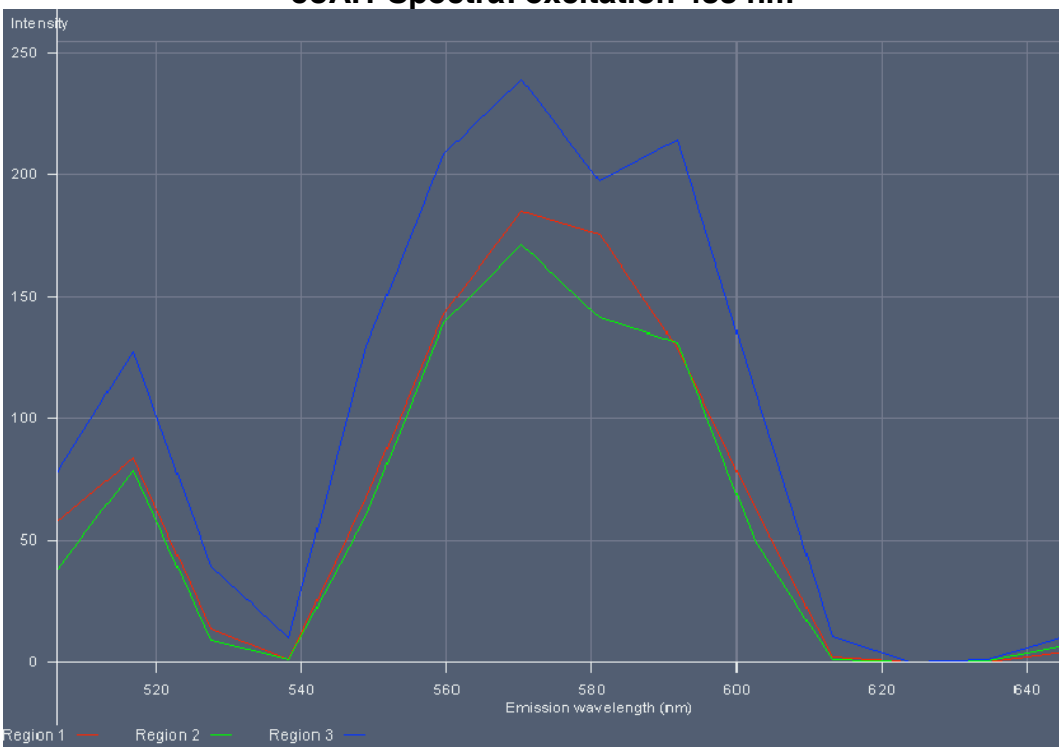



**Figure 5.18. 488 nm excitation of 5oA.1.** A. Curves represent three spectral readings of 5oAi.1 excited at 488 nm taken all at once on the same sample (i.e. replicates). Peaks labeled. Green curve: peak at 517, 570 nm and 592nm. Red curve: peak at 517nm, 570nm and 592nm. Blue Curve: 517 nm, 570 nm, and 592 nm. Emission spectra before 506 nm could not be collected because lower nm would cause a high background signal because 488 nm excitation laser would overlap – false signal. This image has no background signal.

**5oA.1 Spectra: excitation 514 nm**

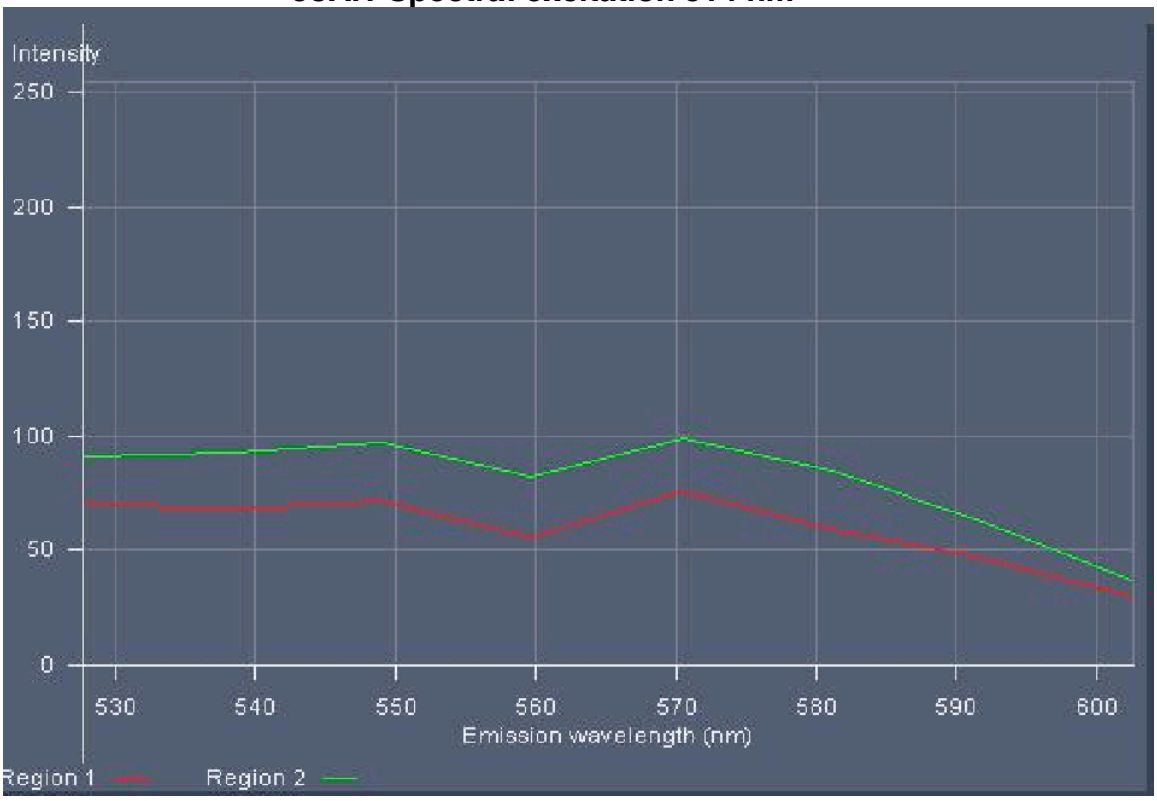



**Figure 5.19. 514 nm excitation of 5oA.1.** A. Curves represent two spectral readings of 5oA.1 excited at 514 nm taken all at once on the same sample (i.e. replicates). Peaks labeled. Green curve: peak at 549nm and 570 nm. Red curve: peak at 549 nm and 570 nm. Emission spectra before 528 nm could not be collected because lower nm would cause a high background signal because 514 nm excitation laser would overlap – false signal. This image has no background signal.

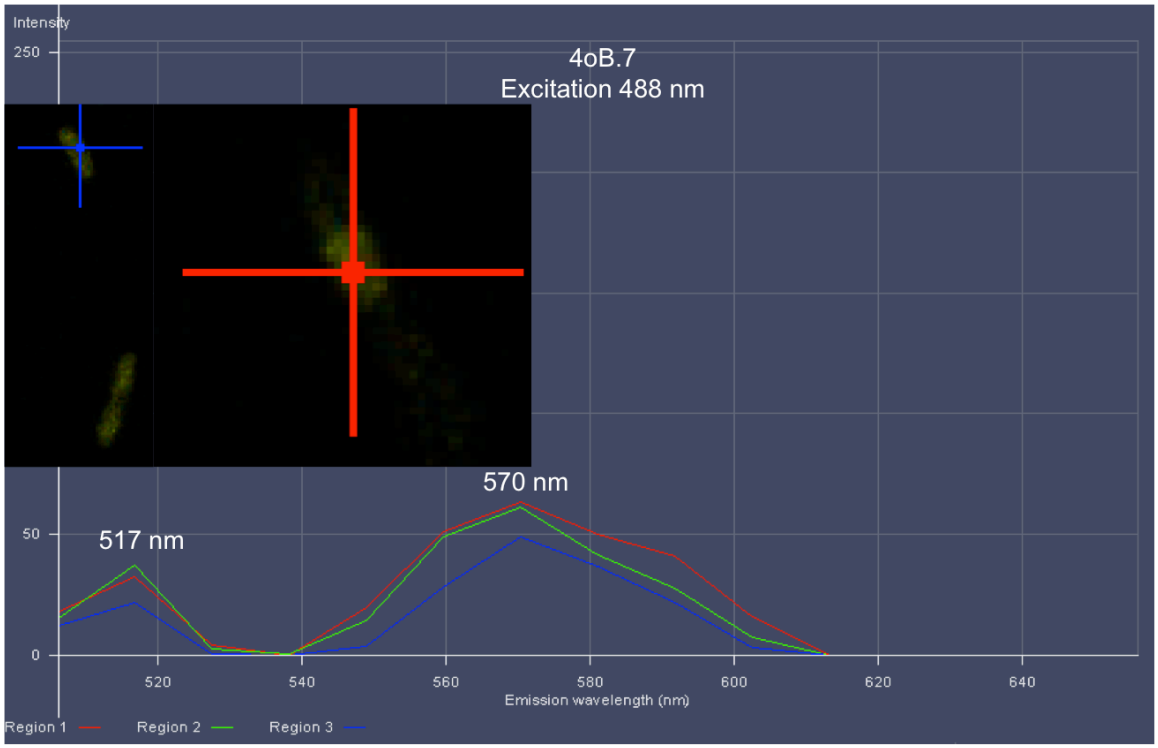
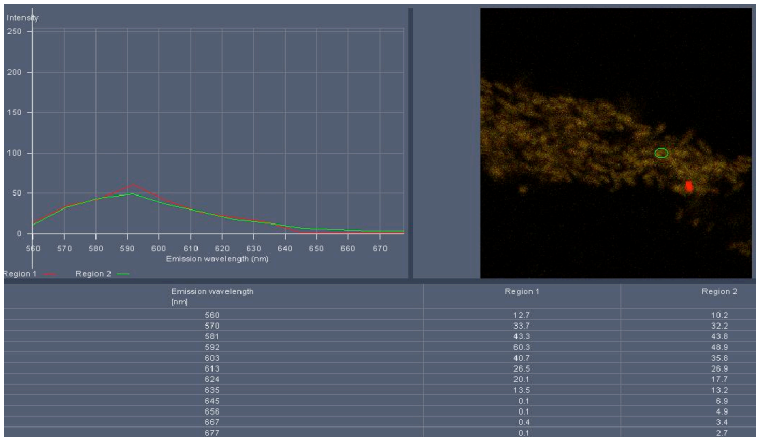



**Figure 5.20. 4oB7 spectra and microscope image.** Curves represent three spectral readings of 4oB.7 excited at 488 nm taken all at once on the same sample (i.e. replicates). Green, red and blue curves all give a minor peak at 517 and a major peak at 570 nm. Images of three bacteria under microscope are depicted in the picture. The bacteria on the right is zoomed in as compared to the two bacteria on the left. Fluorescence was measured at the intersection point of the red, green, and blue crosses (image represents bacteria or section of bacteria selected for spectral reading). Colors of selected areas of measure correspond to the curve color in A. This image has no background signal. Note: COLOR OF BACTERIA DO NOT REPRESENT ACTUAL PHENOTYPE COLOR

**Red and Yellow**

1.5 all red proteins are derived from 1.5 (Figure 5.21 and 5.22). As an example, we will focus in on extant red proteins 4rAi.7 and 4rBi.6, which are closely related (Figure 5.3A). LCA shared between 4rAi.7 and 4rBi.6 is 3r.1 – these two extant proteins have same emission peak at 581 (Figure 5.23). 4rAi.7 has a peak at 581 nm. Its curve is very broad and covers about 80 nm in width and the excitation is observed at 543 nm. Red and green lines are replicates with green line corresponding to selected green circular area in the bacteria image while the red line spectra is gathered from area where the red circle lies – on a bacteria.




| Emission wavelength [nm] | Region 1 | Region 2 |
|---|---|---|
| 560 | 12.7 | 10.2 |
| 570 | 33.7 | 32.2 |
| 581 | 43.3 | 43.8 |
| 592 | 60.3 | 48.9 |
| 603 | 40.7 | 35.8 |
| 613 | 26.5 | 26.9 |
| 624 | 20.1 | 17.7 |
| 635 | 13.5 | 13.2 |
| 645 | 0.1 | 8.9 |
| 656 | 0.1 | 4.9 |
| 667 | 0.4 | 3.4 |
| 677 | 0.1 | 2.7 |

**Figure 5.21. 1.5 ancestor spectra.** Curves represent two spectral readings of 1.5 ancestor excited at 488 nm taken all at once on the same sample (i.e. replicates). Green and red lines give an emission peak at 592nm. Image bacteria under microscope are depicted in the picture. Colors of selected areas of measure correspond to the curve color in A. Bacteria are not bright – protein is not as fluorescent. Emission spectra before 506 nm could not be collected because lower nm would cause a high background signal because 488 nm excitation laser would overlap – false signal. Image has no background signal.

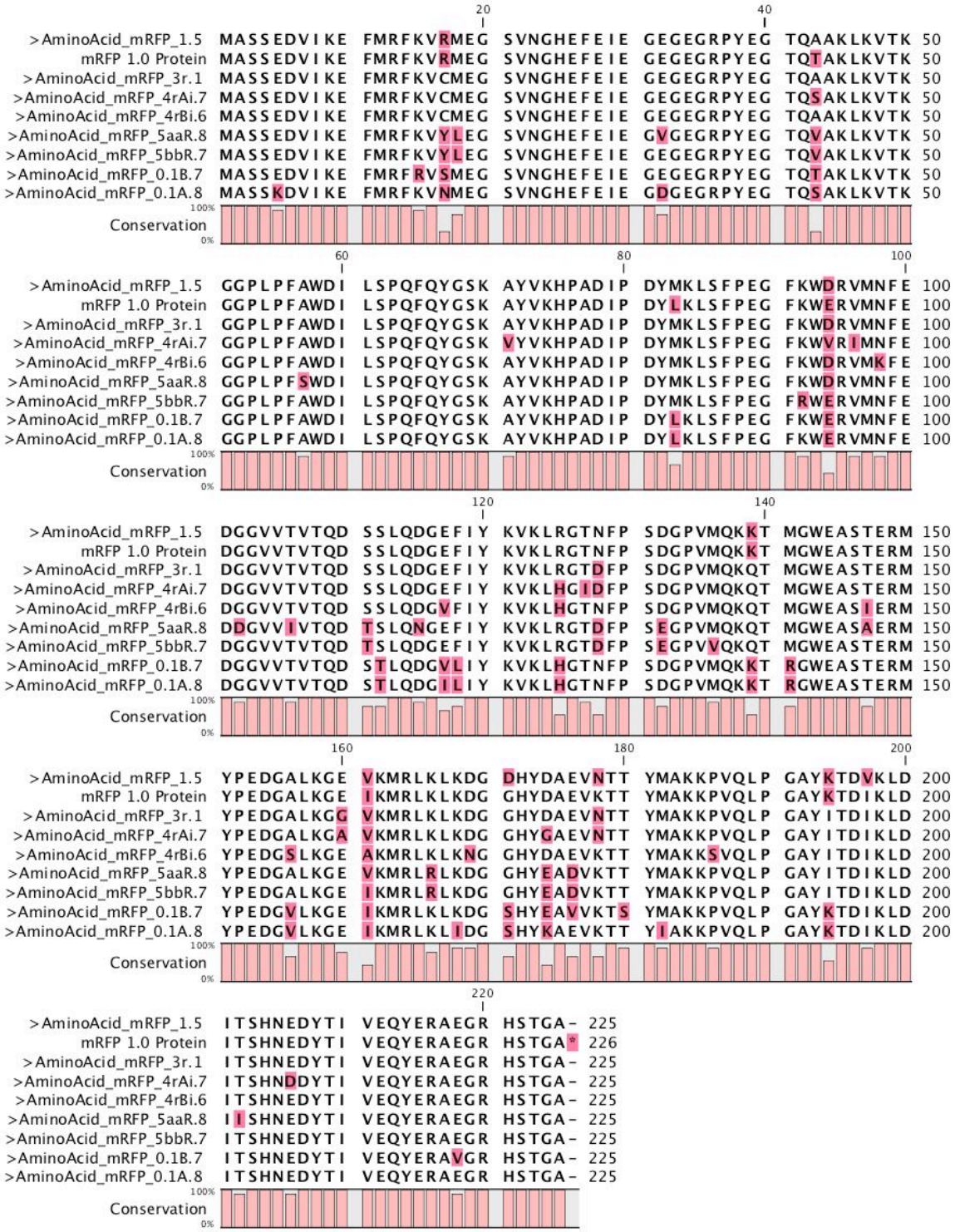
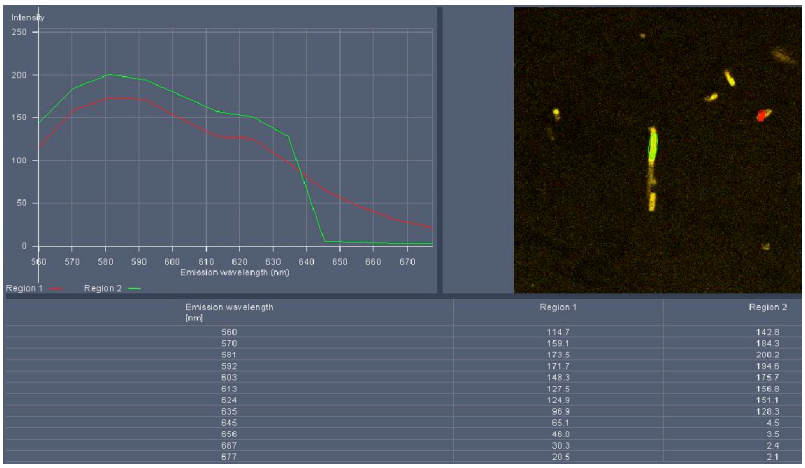



**Figure 5.22. Multiple Sequence alignment of ancestral and extant Red mRFP variant.** All red terminal proteins compared with the 1.5 and 3r.1 ancestor, and with mRFP 1.0.




| Emission wavelength [nm] | Region 1 | Region 2 |
|---|---|---|
| 560 | 114.7 | 142.8 |
| 570 | 159.1 | 184.3 |
| 581 | 173.5 | 200.2 |
| 592 | 171.7 | 194.6 |
| 603 | 148.3 | 175.7 |
| 613 | 127.5 | 156.8 |
| 624 | 124.9 | 151.1 |
| 635 | 96.9 | 126.3 |
| 645 | 65.1 | 4.5 |
| 656 | 46.0 | 3.5 |
| 667 | 30.3 | 2.4 |
| 677 | 20.5 | 2.1 |

**Figure 5.23 4rBi.6 spectra.** 4rBi.6 has a peak at 581 nm. Its curve is also very broad covers about 60 nm in width thus being not as broad as 4rAi.7. Excitation is observed at 543 nm with red and green lines being replicates. Green line corresponds to selected green circular area in the bacteria image and red line spectra was gathered from area where the red circle lies – on a bacteria.

Yellow phenotype was derived only once within the experimental phylogeny, and the details are shown in Figure 5.3A. Seven nonsynonymous mutations occurred from LCA 1.5 until the yellow phenotype occurred in 2.s1. The seven mutations are highlighted in Figure 5.24. It is hard to tell whether or not all seven mutations were necessary for the yellow phenotype based on the mutation site. It looks as though only two of the seven mutations interact with the chromophore environment (Figure 5.24).

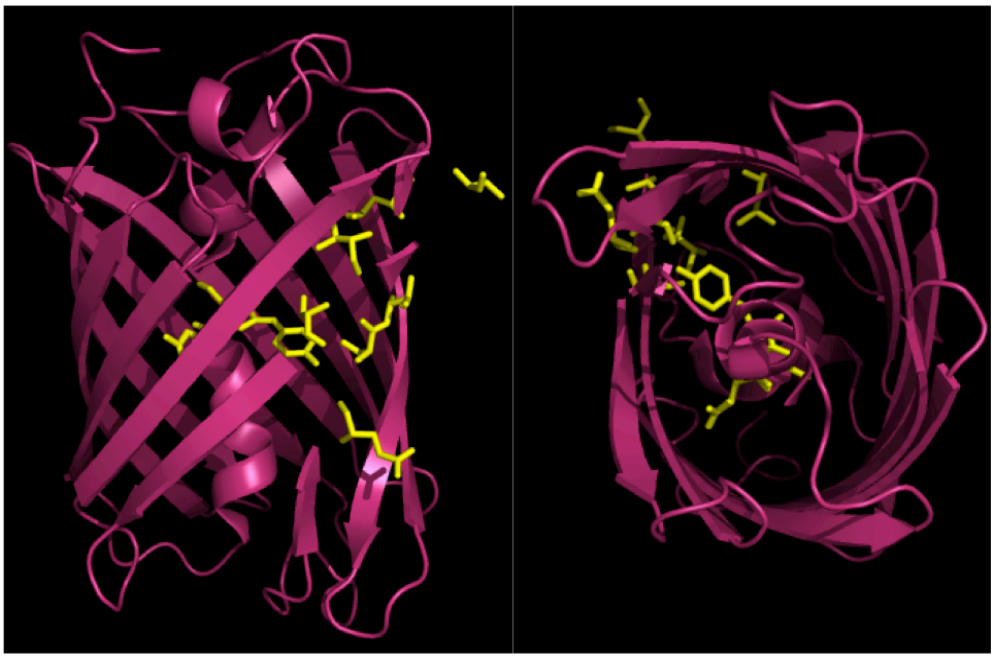


**Figure 5.24 Red 1.5 to yellow 2s.1.** PyMol illustration of the non-synonymous mutations occurring changing the 1.5 mRFP red phenotype to yellow in 2s.1 (Figure 5A.3). Of the major genotypic changes to occur, there are two mutations observed in the chromophore region of the protein which directly influences its phenotype. These mutations are glutamine to methionine at site 66 (Q66M) and tyrosine to tryptophan at site 67 (Y67W).

## Conclusion

In total, the experimental phylogeny offers us a rare opportunity to dissect molecular mechanisms responsible for phenotypic diversification of the evolved fluorescent proteins. These molecular mechanisms include instances of homoplasy (such as convergent or parallel evolution) whereby identical sequences accumulate identical substitutions in a parallel manner that create a new phenotype or whereby dissimilar sequences accumulate identical substitutions in a convergent manner that create a new phenotype, as well as

instances of unique molecular responses that give rise to identical/similar phenotypic properties.

## Biological Realism

Once the phylogeny was constructed, we next questioned whether our sequences evolved in a biologically consistent manner. This is an important question because our phylogeny will ultimately serve to benchmark ancestral sequence reconstruction. In one regard, our evolved sequences are biologically relevant because they are protein-encoded and have experienced functional divergence similar to natural fluorescent proteins. This does not, however, reflect the manner in which natural sequences accumulate mutations under purifying and diversifying selection regimes. More informative descriptors might include comparisons of synonymous and nonsynonymous substitutions among the evolved sequences compared to natural sequences, or comparisons of stationary base frequencies, or comparisons of rate heterogeneity among sites (gamma distribution), etc.

To answer these questions, we have compared our evolved sequences to natural FP sequences across various evolutionary parameters. These parameters are often estimated using specific models of molecular evolution to determine biological consistency. We test such consistency by answering the following questions: Is the model of evolution consistent with real organismal evolution? Is the experimental phylogeny biologically consistent compared with a natural FP phylogeny?

We sought out to find the evolutionary models and parameters that best reflect the experimental phylogeny's evolution in order to understand how the laboratory sequences have evolved. We applied jModelTest to our 19 terminal DNA sequences (from Figure 5.2) and ProtTest to the encoded terminal amino acid sequences in order to determine which evolutionary models and parameters best fit our data using hierarchical Likelihood Ratio Tests (hLRT) and Akaike's Information Criterion (AIC) [2, 3].

The model that best fit the nineteen terminal DNA sequences was the GTR+I model [4] and its jModelTest output is displayed in Table 5.3. The GTR+I is a general, neutral, independent, finite-sites, time reversible model of nucleotide substitution with invariable sites exhibited by known DNA data sets [4-6]. Under the GTR+I model, the rates by which a nucleotide will change to another nucleotide are determined by the six relative rate parameters (a-f) and four nucleotide base frequencies ($\pi_A$, $\pi_C$, $\pi_G$, $\pi_T$) of the given data set. GTR+I also accounts for invariable sites, assuming that rates of substitution vary among sites.

The model that best fit our amino acid sequences was the HIVb+I+G model [7] and its ProtTest output is displayed in Table 5.4.

**Table 5.3. jModelTest Output**. jModelTest selected the GTR+I model as the best-fit model of our 19 terminal DNA FP sequences. The reverse relative rate values equal their reciprocal. Ti – transition; Ts - transversion

| | DNA Substitution Type | | GTR+I #1 |
|---|---|---|---|
| | | partition | 012345 |
| | | -lnL | 3360.6872 |
| | | K | 45 |
| Adenine Frequency | | freqA = | 0.2964 |
| Cysteine Frequency | | freqC = | 0.2221 |
| Guanidine Frequency | | freqG = | 0.2468 |
| Thiamine Frequency | | freqT = | 0.2348 |
| Relative Rate A→C | Ts | R(a) [AC] | 0.4864 |
| Relative Rate A→G | Ti | R(b) [AG] | 2.5498 |
| Relative Rate A→T | Ts | R(c) [AT] | 2.0143 |
| Relative Rate C→G | Ts | R(d) [CG] | 0.3222 |
| Relative Rate C→T | Ti | R(e) [CT] | 4.0405 |
| Relative Rate G→T | Ts | R(f) [GT] | 1.0000 |
| | | p-inv -+I | 0.5080 |

**Table 5.4. ProtTest Output**. ProtTest selected the HIVb+I+G model as the best-fit model of our 19 terminal amino acid FP sequences.

| | HIVb+I+G #1 |
|---|---|
| parameters | 37 (2 + 35 branch length estimates) |
| Gamma shape (4 rate categories) | 1.124 |
| proportion of invariable sites | 0.377 |
| -lnL | 1865.27 |

Both GTR+I and HIVb+I+G represent sequence substitution patterns of real organismal sequences over time. Our parameter estimates are consistent with parameters estimates of biological sequences based on our past experiences [9]. As these models are representative of our experimental phylogeny's sequence evolution, we therefore conclude that the laboratory phylogeny has evolved in a manner consistent with biological evolution and we therefore conclude that, to date, we have achieved a substantial amount of sequence and phenotypic diversity from a single red fluorescent protein.

The functional divergence experienced by the experimental phylogeny was next assessed for biological relevance. We can physically witness the functional divergence within the mRFP experimental phylogeny because we can see the multiple phenotypes that have arisen from the single ancestral red phenotype. However, this phylogeny was artificially created in the laboratory. As such, we set out to determine whether this artificial functional divergence is consistent with real biological functional divergence.

Evaluating the biogenicity of the experimental tree's functional divergence was tested by comparing the parameter estimates associated with the evolved mRFP phylogeny and 19 randomly selected natural coral fluorescent genes of all known fluorescent coral phenotypes: cyan, green, and red. The 19 natural coral FPs alignment is referred to as GFP19 (see table 5.5 for GFP 19 details). Parameters were estimated via PAMLv. 4.5.

**Table 5.5. Identity of nineteen natural fluorescent sequences.** Coral genes used in assessing the biological revelance of the laboratory phylogeny. Colors (Cyan, green and red) represent the species spectral color class.

| | Protein ID | Genbank | Taxonomy |
|---|---|---|---|
| 1 | cFP484 | AF168424 | Clavularia sp. (Anthozoa , Alcyonaria, Alcyonacea) |
| 2 | mcRFP | AY362545 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 3 | mcFP | AY056460 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 4 | mc5 | AY181556 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 5 | monfavGFP2 | AF401282 | Montastraea faveolata (Anthozoa, Hexacorallia, |
| 6 | mcavGFP | AF406766 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 7 | mcavFP_6 | AY037769 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 8 | mc3 | AY181554 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 9 | mc4 | AY181555 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 10 | mc6 | AY181557 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 11 | M. cavernosa | AF384683 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 12 | scubGFP1 | AY037767 | Scolymia cubensis (Anthozoa, Hexacorallia, Scleractinia) |
| 13 | dendGFP | AF420591 | Dendronephtya sp. (Anthozoa, Alcyonaria, Alcyonacea) |
| 14 | dis3GFP | AF420593 | Discosoma sp. 3 (Anthozoa, Zoantharia, |
| 15 | Synthetic? | AY218848 | Derived from Montastraea cavernosa |
| 16 | Kaede | AB085641 | Trachyphyllia geoffroyi (Anthozoa, Hexacorallia, |
| 17 | mcavFP_7.5 | AY037770 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 18 | mc1 | AY181552 | Montastraea cavernosa (Anthozoa, Hexacorallia, |
| 19 | rfloRFP | AY037773 | Ricordea florida (Anthozoa, Zoantharia, |

Parameter estimates of base frequency in codon positions 1, 2 and 3 are all very similar between the laboratory evolved phylogeny and the natural FP phylogeny (Table 5.6). The tree length is the sum of all the branch lengths, a branch length being the number of nucleotide substitutions per codon. The tree length of the natural FP phylogeny is about three times the tree length of our current phylogeny, however, this will be addressed accordingly as we continue to build our phylogeny through extension of our terminal branches and thus increasing our overall tree length.

**Table 5.6 Parameter estimates of natural versus experimental fluorescent proteins.** Parameter estimates associated with the laboratory evolved mRFP sequences in comparison to nineteen randomly selected natural fluorescent genes from Anthozoa coral species having all known fluorescent coral phenotypes: cyan, green, and red. Parameters, including branch lengths, were estimated in PAML [9].

| Parameter | Our evolved mRFPs to date | Natural Fluorescent Genes |
|---|---|---|
| Base Freqs Codon Position: 1 | T:0.168 C:0.194 A:0.258 G:0.380 | T:0.178 C:0.170 A:0.308 G:0.345 |
| Base Freqs Codon Position: 2 | T:0.257 C:0.212 A:0.367 G:0.163 | T:0.291 C:0.162 A:0.369 G:0.177 |
| Base Freqs Codon Position: 3 | T:0.263 C:0.280 A:0.219 G:0.238 | T:0.263 C:0.262 A:0.209 G:0.266 |
| Base Freqs Averages | T:0.229 C:0.229 A:0.281 G:0.260 | T:0.244 C:0.198 A:0.295 G:0.263 |
| Tree Length | 2.302 | 6.782 |
| Tree length nonsynon rate (dN) | 0.434 | 1.329 |
| Tree length for synon rate (dS) | 1.653 | 5.064 |
| Omega dN/dS ratio | 0.262 | 0.262 |
| Kappa (ts/tv) | 2.39 | 1.87 |
| Alpha (gamma, K=8) | 1.68 | 2.48 |

## Conclusion

This analysis demonstrates that our distribution of substitutions in the evolved phylogeny is analogous to the substitution patterns of natural fluorescent genes. This sequence diversity was evolved in a manner consistent with biological evolution (albeit in our case artificially evolved). Since our FP phylogeny is incomplete, we have the ability to match the natural FP phylogeny effectively through building our phylogeny in a manner that is biologically consistent with natural FP evolution.

## References


1. Sample, V., R.H. Newman, and J. Zhang, *The structure and function of fluorescent proteins.* Chem Soc Rev, 2009. **38**(10): p. 2852-2864.
2. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21**(9): p. 2104-5.
3. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.
4. Tavare, S., *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.* American Mathmatical Society, 1986. **17**: p. 57-86.
5. Rodriguez, F., et al., *The General Stocastic Model of Nucleotide Substitution.* J. ther. Biol., 1990. **142**: p. 485-501.
6. Lanave, C., et al., *A new method for calculating evolutionary substitution rates.* J Mol Evol, 1984. **20**(1): p. 86-93.

7. David C. Nickle, L.H., Mark A. Jensen, Peter B. Gilbert, James I. Mullins, Sergei L. Kosakovsky Pond, *HIV-Specific Probabilistic Models of Protein Evolution.* PLoS ONE, 2007. **2**(6): p. e503.

8. Gaucher, E.A., M.M. Miyamoto, and S.A. Benner, *Function–structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors.* PNAS, 2001. **98**(2): p. 548-552.

9. Yang, Z., *PAML 4: Phylogenetic Analysis by Maximum Likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.

# CHAPTER 6

# ANCESTRAL SEQUENCE RECONSTRUCTION

In this preliminary study, the generated leaf/tip sequences from the evolved phylogeny have been used to estimate ancestral genotypes and phenotypes. Since the leaf/tip sequences have been sequentially evolved from nodes on the experimental phylogeny in the laboratory, we know the true ancestral genotypes and phenotypes. Thus, fluorescent protein experimental phylogeny has presented us with the unique opportunity to compare/contrast different approaches attempting to reconstruct ancestral sequences from biologically relevant conditions. Although similar benchmarking was achieved for phylogeny-building algorithms, reconstruction of restriction sites and growth rates by Hillis & Bull and others [1-5], our proposed work represents the first time evolved sequences have been used to benchmark ancestral sequence reconstruction approaches to address issues of ambiguity and bias associated with both reconstructed genotypes and phenotypes. The following section provides an understanding as to how this phylogeny will be applied in benchmarking ASR methods. As an example, node 3s.1 from the experimental phylogeny was selected to conduct a preliminary ASR analysis in order to demonstrate what can be determined with a known, experimental phylogeny.

## Phylogenetic Tree Construction

A tree topology is needed for input in order to perform ancestral sequence reconstruction (ASR). We elected to use three tree topologies in our ASR analyses in order to determine how different topologies affect ancestral sequence reconstruction. The nineteen terminal DNA sequences were used as input to infer a phylogeny using MrBayes v.3.2.1 run for twenty million generations under the best-fit general-time reversible plus gamma distribution (GTR+I) model according to jModelTest [6] (other models were also tested but they did not influence the topology of the phylogeny). We refer to this MrBayes DNA tree topology as MB_DNA (Figure 6.1). Then, the nineteen terminal amino acid sequences were used to infer a phylogeny using MrBayes run for ten million generations under the JTT+I+G model. We refer to this MrBayes amino acid tree topology as MB_AA (Figure 6.2) In addition to these two trees, the true tree topology of the evolved laboratory FPs was also tested in our ASR analyses because neither the DNA nor amino acid analyses using MrBayes generated the true tree topology (Figure 6.3).

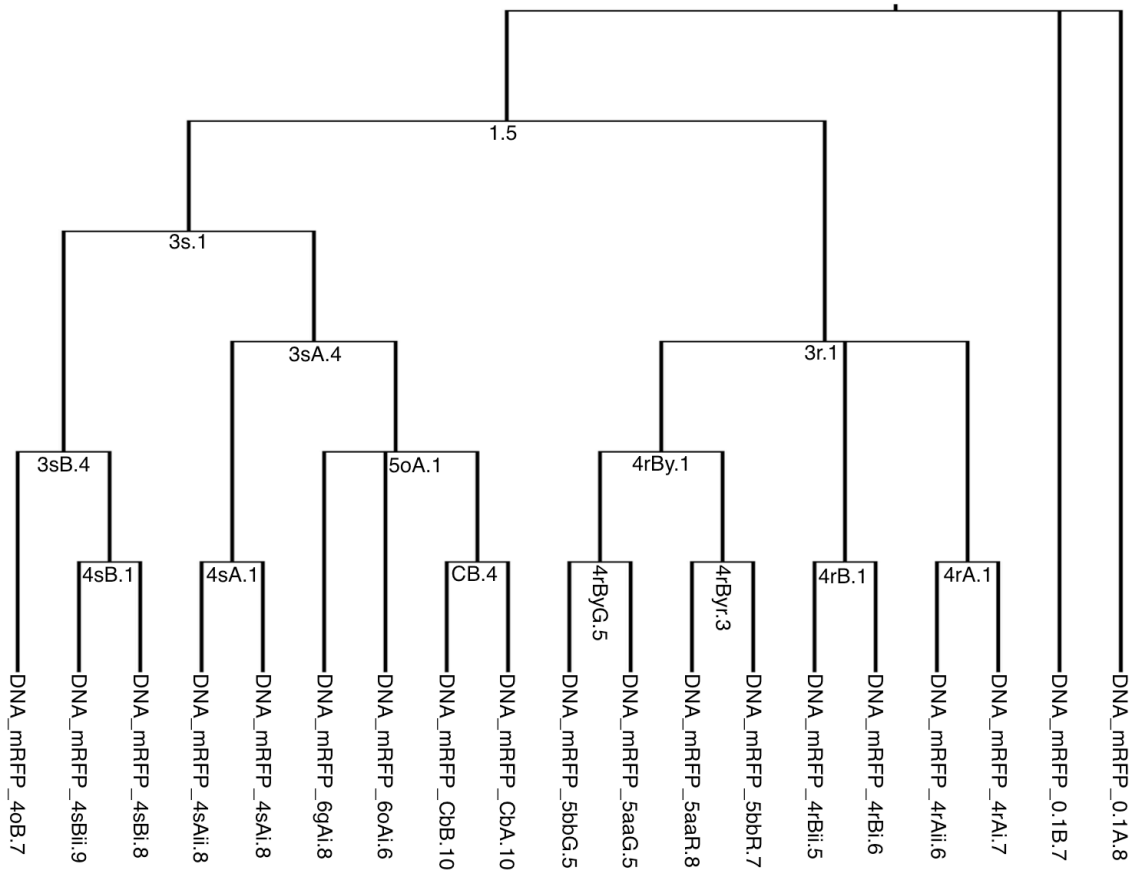



**Figure 6.1. MB_DNA cladegram.** Tree topology of our evolved FP tree inferred from MrBayes using the nineteen terminal DNA sequences.

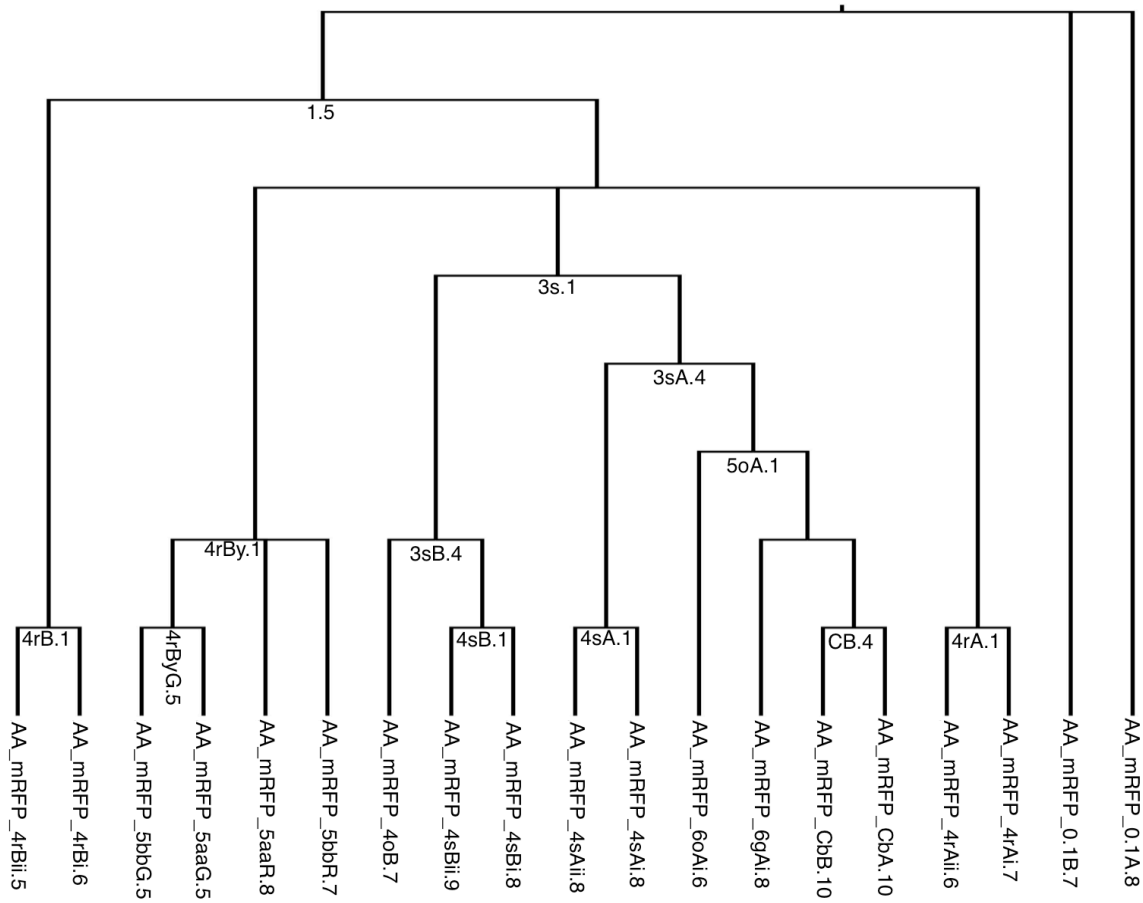



**Figure 6.2. MB_AA cladegram.** Tree topology of our evolved FP tree inferred from MrBayes using the nineteen terminal amino acid sequences.

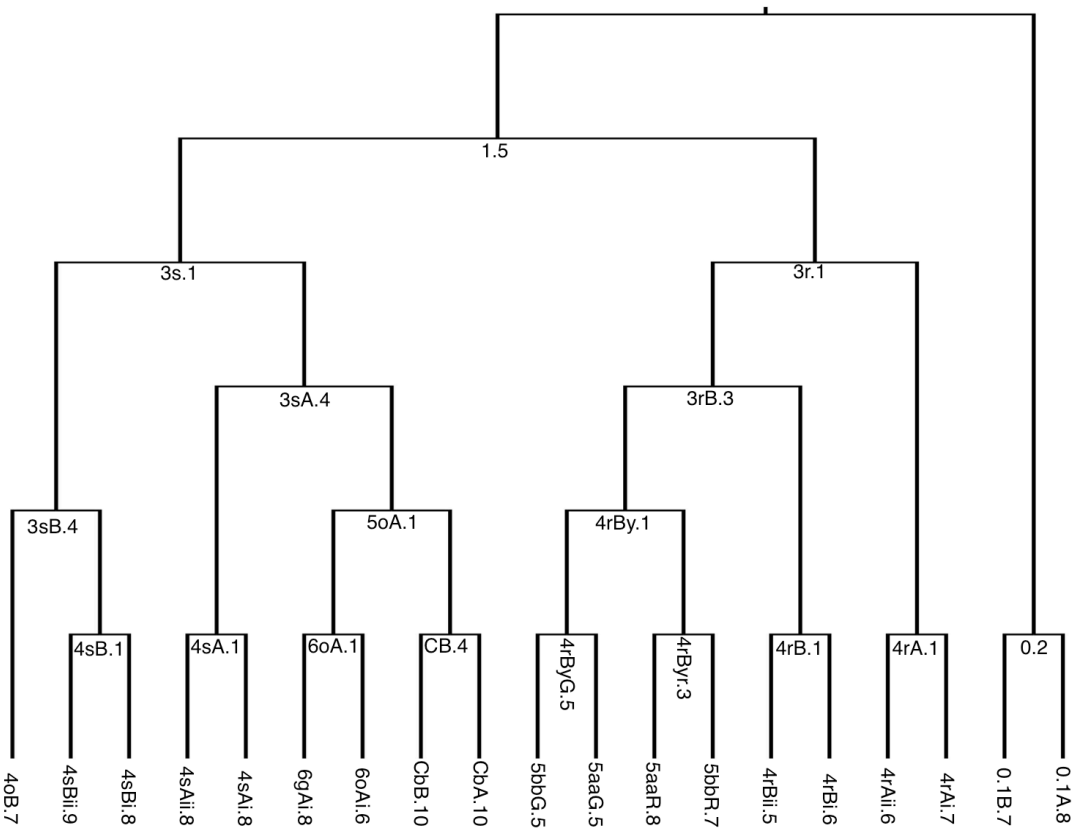



**Figure 6.3. TT cladegram.** This tree represents the true topology of the evolved FP tree, topology built in PAML v. 4.1.

## Comparison of Tree Topologies

The best trees from the MrBayes analyses (MB_DNA and MB_AA) are only partially resolved due to instances of polytomies. These polytomies arise from very short internal branches and are incorrect because the true topology (TT) did not contain any polytomies. The inference that is most similar to the TT is the MB_DNA tree. This tree has missing nodes, namely, 0.2, 3rB.3 and 6oA.1. Node 0.2 is missing as a result of a polytomy. In the true tree, node 1.5 is rooted

by the outgroup branch which connects node 1.5 to node 0.2, and then bifurcates at node 0.2 into the 0.1Ai.8 and 0.1B.7 terminal taxa. The polytomy resulting in the absence of 0.2 is misleading because it leads one to think that node 1.5 is rooted by two distant outgroups. Both nodes 3rB.3 and 6oA.1 are also missing on the MB_DNA tree; instead, the nodes are replaced by polytomies.

The MB_AA is missing nodes 3r.1, 0.2, 4rBy.3 and 6oA.1. 3r.1's absence significantly confounds the true relationships within the FP phylogeny since this node is so deeply rooted within the phylogeny. Though both the DNA and AA trees lack node 6oA.1, they do not agree as to how it's missing. Instead of node 6oA.1, the MB_AA tree shows an unknown node in the place where 6oA.1's should be. This unidentified node separates the true sister taxa 6oAi.6 and 6gAi.8 into more anciently derived roots. In the true tree topology the very short internal branch connecting 5oA.1 to 6oA.1 is highly confounding in phylogenetic inferences, resulting in trees containing polytomies or incorrect branching patterns.

**Topologies used for Ancestral Sequence Reconstruction**

DNA and amino acid sequences were used as input to infer ancestral sequences using the true topology. We refer to the output of these analyses as TT_DNA and TT_AA. Alternatively, the constructed phylogeny using DNA in MrBayes was used as input to infer ancestral DNA sequences only (MB_DNA). Further, the constructed phylogeny using amino acid in MrBayes was used as

input to infer ancestral amino acid sequences only (MB_AA). Thus, a total of four different ASR analyses were conducted on three different topologies (Figure 6.4):

1. MrBayes Topology using DNA = MB_DNA

2. MrBayes Topology using amino acid = MB_AA

3. True Tree Topology using DNA = TT_DNA

4. True Tree Topology using amino acid = TT_AA

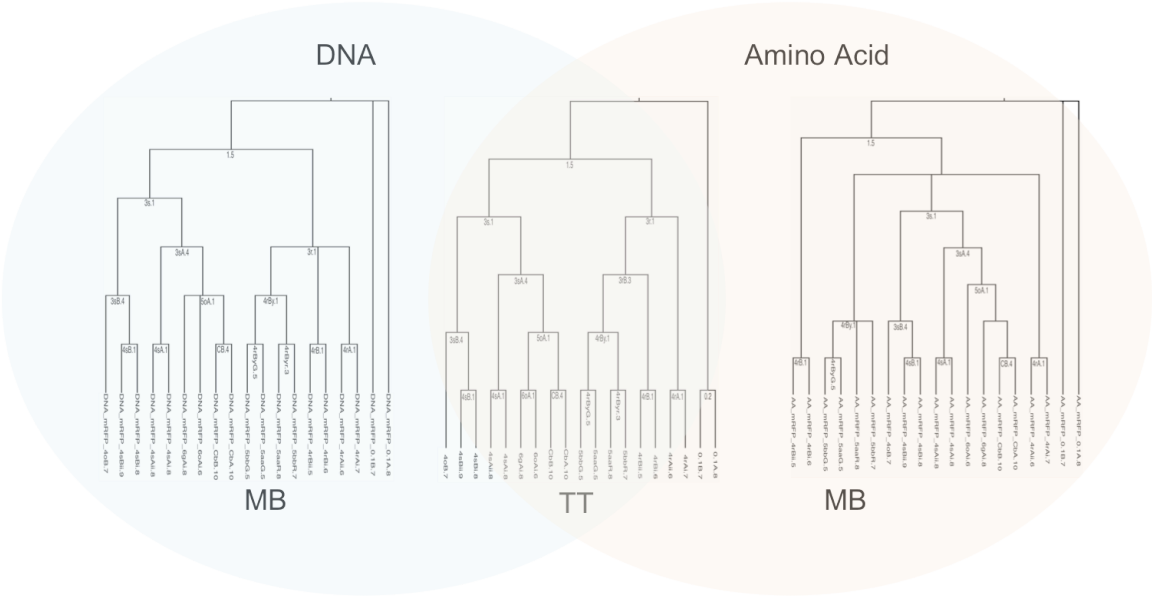



**Figure 6.4. ASR trees.** Two topologies, TT_DNA and MB_DNA, were used to infer ancestral DNA sequences and two topologies (TT_AA and MB_AA) were used to infer ancestral amino acid sequences. The true topology (TT) is identical between the DNA and amino acid analyses.

## Sequence Reconstruction

Given that the experimental phylogeny is still in the process of being built, there is no use of performing detailed ASR analyses on all nodes of the phylogeny at this time. That said, we chose to perform preliminary ASR tests on node 3s.1 in order to obtain a greater understanding of what lies ahead for us. We chose this particular node because 1) it is deep within the tree and 2) this ancestor has given rise to a large amount of phenotypic diversity. The intention is that this preliminary analysis provides a glimpse into how ancestral sequence reconstruction can be benchmarked. Several different ASR analyses were conducted to infer the ancestral sequence at node 3s.1. These analyses included different inference methods (parsimony vs. likelihood) and different models of evolution (rate heterogeneity, transitions/transversions, base frequencies).

This section will begin by describing the maximum parsimony results and then move on to the maximum likelihood results. The section will then end with comparisons of the different results.

### Maximum Parsimony

Inferred MP sequences are denoted by adding '_MP' to the name of the topology the sequence was inferred from:

3s.1 inferred sequence using TT_DNA topology = TT_DNA_MP

3s.1 inferred sequence using MB_DNA topology = MB_DNA_MP

3s.1 inferred sequence using TT_AA topology = TT_AA_MP

3s.1 inference using MB_AA topology = MB_AA_MP

*Results and Discussion*

The TT_DNA_MP and MB_DNA_MP sequences were translated to amino acid sequences so that all four MP sequences could be compared in a multiple sequence alignment (Figure 6.5). The inferred TT_AA_MP and MB_AA_MP amino acid sequences are identical in the ASR analyses. For simplicity, these sequences are referred as AA_MP throughout since the topology is irrelevant for the ASR inference. The translated TT_DNA_MP and MB_DNA_MP sequences, however, are different from each other and from AA_MP. Leaving us with three unique MP ancestral sequences (Table 6.1 and Figure 6.6).

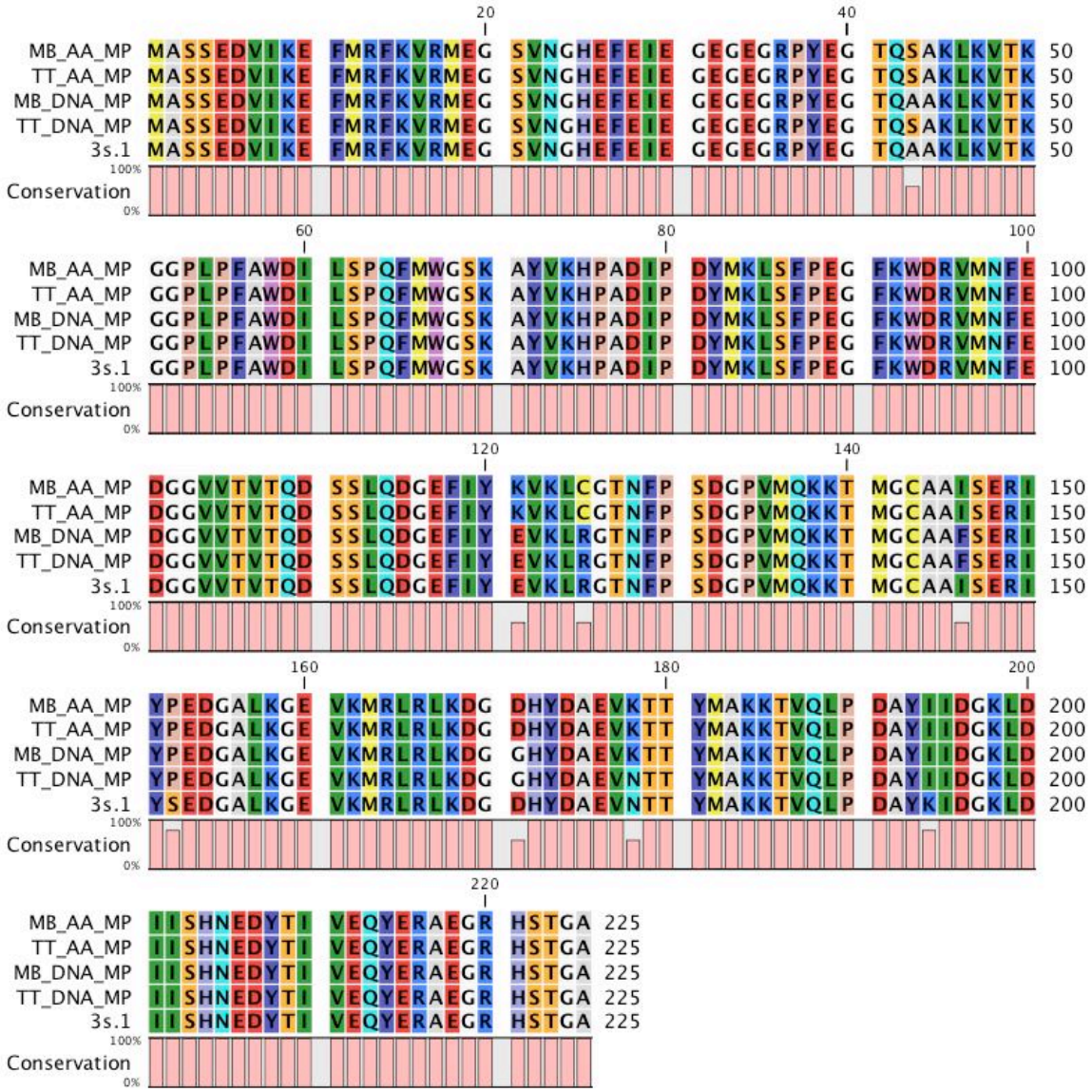



**Figure 6.5. Multiple Sequence Alignment of MP inferences and 3s.1.** Multiple sequence alignment (MSA) of the translated MB_DNA_MP nucleotide sequence, the translated TT_DNA_MP nucleotide sequence, the MB_AA_MP amino acid sequence, the TT_AA_MP amino acid sequence, and the 3s.1 amino acid sequence. Amino acid sequences aligned using CLC bio v. 4.1.2.

**Table 6.1. Maximum parsimony inference comparison.** The number of incorrectly inferred residues for each unique MP sequence is listed. The specific residue replacement is listed as well.

| Inferred Sequence Name | # Sites Wrong | Incorrectly Inferred residues |
|---|---|---|
| TT_AA_MP | 6 | A43S; E121K; R125C; S152P; N178K; |
| TT_DNA_MP | 5 | A43S; I146F; S152P; D171G; K194I |
| MB_DNA_MP | 5 | I146F; S152P; D171G; N178K; K194I |

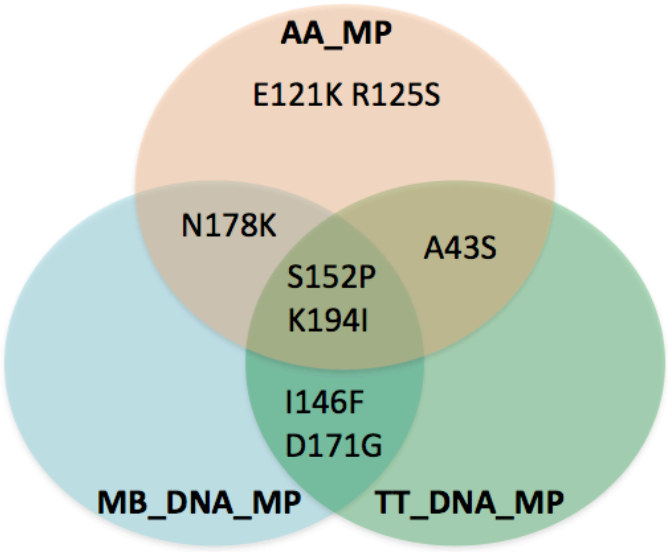



**Figure 6.6. Overlapping incorrectly inferred residues by maximum parsimony.** Inferred MP sequences listing their incorrectly inferred residues shown in a Venn diagram. AA_MP has a total of 6 incorrectly inferred residues, TT_DNA_MP has a total of 5 incorrectly inferred residues, MB_DNA_MP has a total of 5 incorrectly inferred residues.

All ancestral MP inferred sequence have two incorrectly inferred residues in common: S152P and K194I. Site 152 has been replaced within the experimental phylogeny a total of three times (Figure 4.3a). In the branch connecting node 1.5 to node 3s.1, a proline is replaced by a serine at 152. Shortly there after, site 152 is reverted back to a proline along the branch leading to node 3sA.4, and along the branch leading to 4sB.1. All of 3sA.4's and 4oA.1's descendants (8 taxa) have a proline at site 152, thus, of 3s.1's nine total modern descendants, 8 taxa have a proline at residue 152. In addition, the other modern descendants not from direct descent of 3s.1 also have a proline at site 152. Thus, the pattern that is seen from the extant sequences indicates that the eighteen sequences have a proline at site 152 and only one sequence, 4oB.7, has a serine at site 152. The most parsimonious solution would be to assume site 152 only changed once from proline to a serine along the terminal branch leading to 4oB.7, as opposed to three nonsynonymous 152 substitutions, which were actually observed.

Site 194 was replaced to an isoleucine in the branch connecting 1.5 to 3r.1. No further nonsynonymous mutations at this site were observed in any of the 3r.1's descendants leading to all of 3r.1's eight terminal taxa retaining the isoleucine.

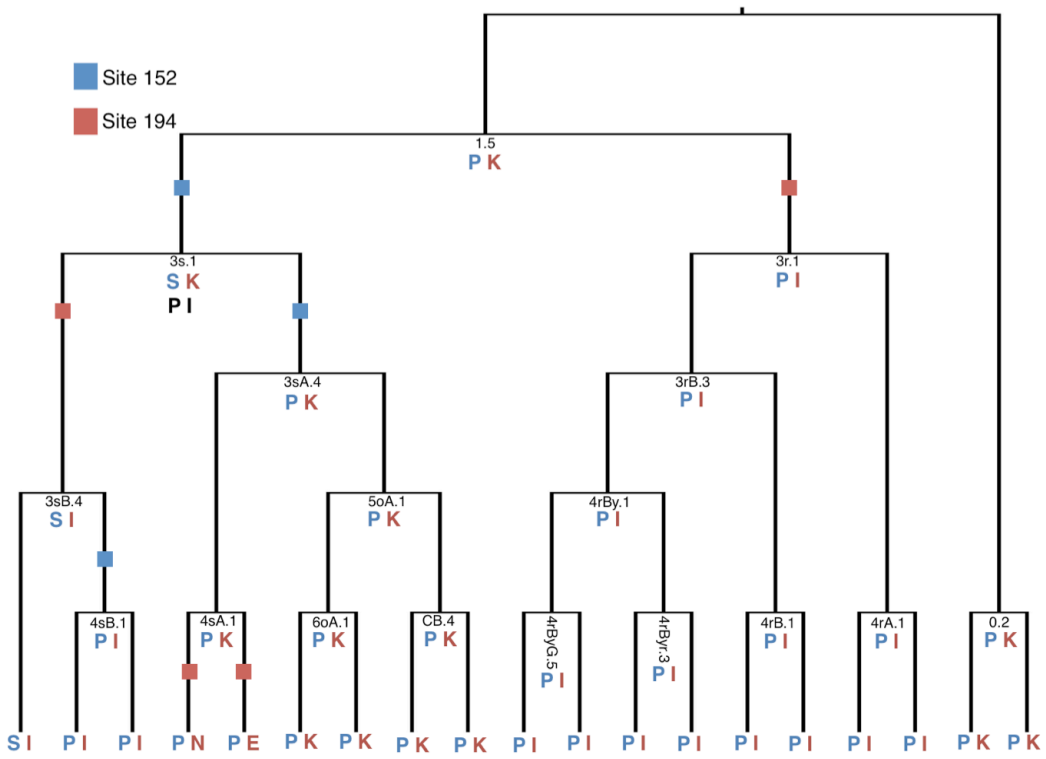



**Figure 6.7. Incorrectly inferred sites 152 and 194.** The figure represents the residues present on Site 152 and Site 194. Residues at Site 152 are indicated in Blue and residues at site 194 are indicated in red. Based on the extant sequences, residues at site 152 and 194 for node 3s.1 were inferred to be proline and isoleucine respectively by MP. This is an incorrect inference as shown by the squares located on the branches which indicate an amino acid replacement, with blue indicating site 152 and red, site 194.

TT_DNA_MP and MB_DNA_MP have two mutations in common: I146F and D171G. We have seen the I146F mutation before in the color change event leading from yellow to orange phenotype as mentioned in the previous section. This wrong inference makes sense since nine out of the ten total terminal branches have a phenylalanine at residue 146.

Based on MP sequence reconstruction alone, the true tree and MrBayes topologies performed identically. The performance of the TT_AA and MB_AA topology was identical in reconstruction of the 3s.1 sequence. The TT_DNA topology incorrectly inferred 5 residues and the MB_DNA incorrectly inferred 5 residues too. Even though these sequences mismatched at one site, we can say that the topologies performed identical in regards to reconstructing the sequence most similar to the true ancestor. By further resurrection of the inferred ancestral functions we may be able to deduce which topology inferred the more accurate sequence. The use of DNA sequence input versus amino acid sequence input did make a difference. DNA topologies used with DNA input are more accurate at inferring the ancestral sequence of 3s.1 over amino acid inferred topologies and amino acid sequence data.

**Maximum Likelihood**

Models of evolution are required to perform ML inferences. To test for discrepancies among the models of evolution used for sequence reconstruction in the ML method, multiple models were chosen.

Models for DNA input using TT_DNA and MB_DNA topologies.

The model that best fit our DNA sequences was GTR + G under jModelTest (ranked #13). To test the effects of the gamma distribution we also used the GTR model (jModelTest ranked best-fit model # 61). We also wanted to compare the performance of the better fit more complex models to less fit, simpler models to see how using less parameters will affect sequence

reconstruction, thus the Jukes Cantor (jModelTest ranked best-fit model # 88) and Kimura's 1980 model (jModelTest ranked best-fit model # 80) were used.

Summary:

Four reconstructions for each DNA topology, eight total reconstructions:

TT_DNA_GTR+G

TT_DNA_GTR

TT_DNA_JC

TT_DNA_K80

MB_DNA_GTR+G

MB_DNA_GTR

MB_DNA_JC

MB_DNA_K80

Models for AA input using TT_AA and MB_AA topologies.

The best-fit model for our amino acid sequences that was available to infer ancient sequences in PAML [7] was the JTT + G model predicted by ProtTest (ranked best-fit model #8). To test the effects of gamma we also used the JTT model (Protest ranked best-fit model # 41). We also wanted to compare the performance of these better-fit, more complex models to less-fit, simpler models to see how using less parameters would affect sequence reconstruction, thus the DAYHOFF (Protest ranked best-fit model # 86) model was used.

Three reconstructions for each AA topology, six total reconstructions:

TT_AA_JTT+G

TT_AA_JTT

TT_AA_DAYHOFF

MB_AA_JTT+G

MB_AA_JTT

MB_AA_DAYHOFF

Total ML models used with DNA topologies for DNA input and with AA topologies for AA input total 14 sequences

Results and Discussion

Out of the fourteen 3s.1 ML ancestral inferences only ten were physically resurrected in the laboratory. GTR and JC model for DNA sequences were not used, therefore all sequences except for TT_DNA_GTR, TT_DNA_JC, MB_DNA_GTR, and MB_DNA_JC are discussed below.

Out of all ML DNA inference methods, there are two instances when the tree topology used had no affect on the 3s.1 ancestral inference under the GTR+G and K80 nucleotide substitution models. TT_DNA_ML_GTR+G and MB_DNA_ML_GTR+G have an identical sequence and this sequence will now be referred to as DNA_ML_GTR+G. TT_DNA_ML_K80 and MB_DNA_ML_K80 also have an identical sequence. This sequence will now be referred to as DNA_ML_K80. DNA_ML_GTR+G had 5 incorrectly inferred residues: E121K, I146F, S152P, D171G, and K194I when compared with the true 3s.1 ancestor (Figure 6.8). DNA_ML_K80 had 6 incorrectly inferred residues: E121K, I146F, S152P, D171G, N178K and K194I when compared with the true 3s.1 ancestor

(Figure 6.8). Thus, we only have two DNA inference sequences (out of a possible four) to work with in our resurrection studies. DNA_ML_GTR+G and DNA_ML_K80 all share the same incorrect inferences, except K80 has an extra incorrect site: N178K. In this case, the best-fit GTR+G model out-performs the K80 model in inferring the more accurate ancestral sequence since the GTR+G had fewer incorrectly inferred residues that the K80 model. This also means that in our ML_DNA ASR analyses the topology was not an important factor in assessing the more accurate 3s.1 inference, but the model of nucleotide substitution mattered.

We next analyzed the ML amino acid inferences. MB_AA_ML_JTT+G and MB_AA_ML_JTT have an identical sequence. This sequence will now be referred to as MB_AA_ML_JTT+/-G. MB_AA_ML_JTT+/-G had 5 incorrectly inferred residues: E121K, S152P, D174E, N178K and K194I (Figure 6.9). This means that the gamma distribution was not an important factor for determining the 3s.1 inference under the JTT model. No other ancestral inferences were identical, so we our analyses included a total of five ML_AA 3s.1 ancestral inferences.

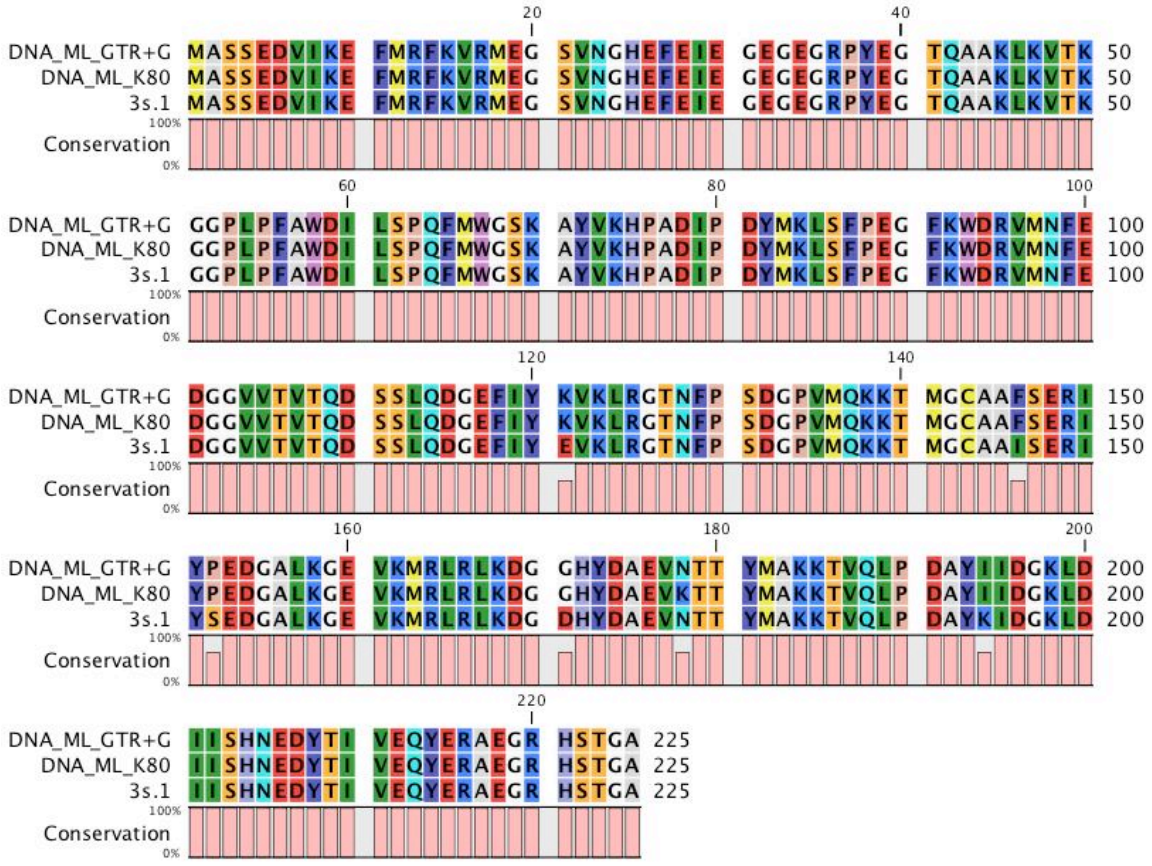



**Figure 6.8. Multiple sequence alignment of 3s.1 and ML DNA inferences.** Multiple sequence alignment of the translated ML DNA inferences DNA_ML_K80 and DNA_ML_GTR+G. DNA_ML_GTR+G had incorrectly inferred residues and DNA_ML_K80 had 8 incorrectly inferred residues. Amino acid sequences aligned using CLC bio v. 4.1.2.

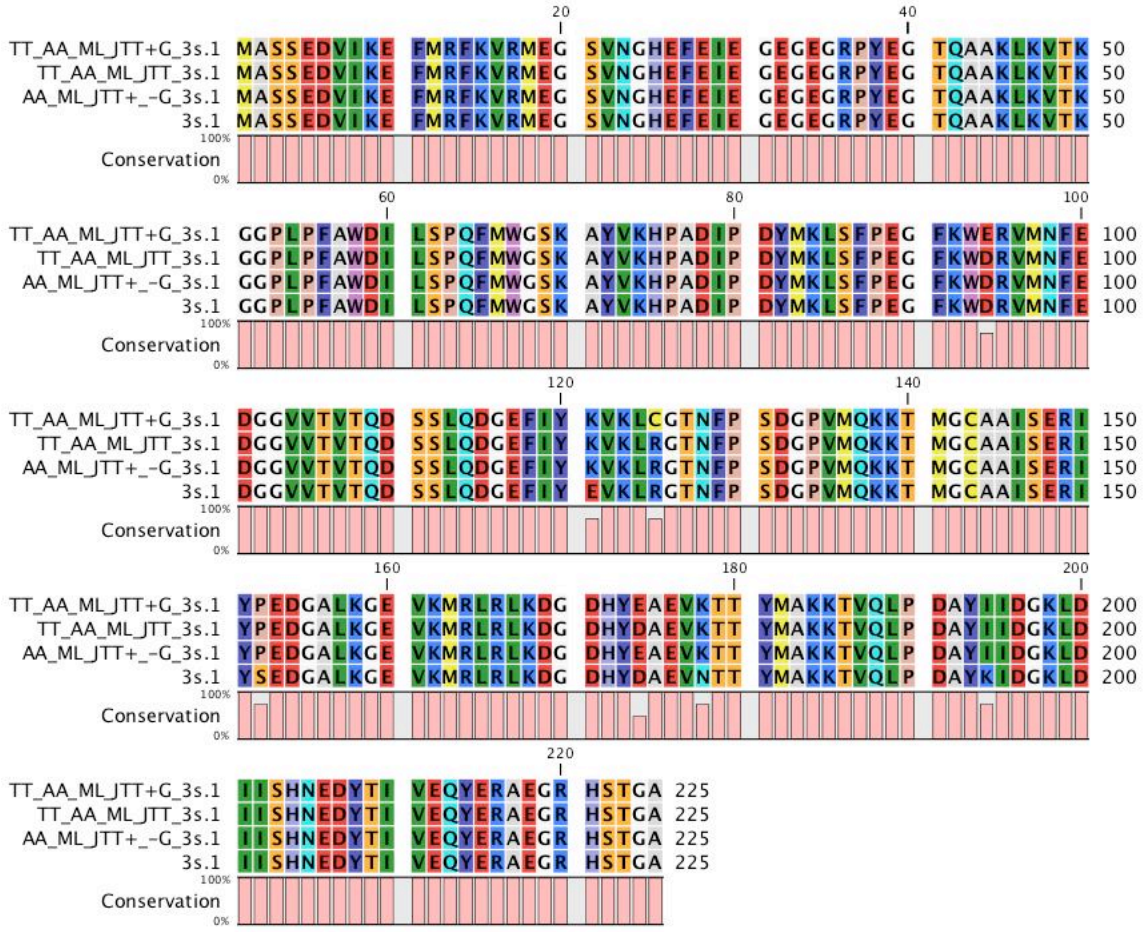



**Figure 6.9. Multiple sequence alignment of 3s.1, JTT, and JTT+G amino acid inferences.** Multiple sequence alignment of AA_ML_JTT+/-G, TT_AA_ML_JTT, and TT_AA_ML_JTT+G. AA_ML_JTT+/-G had 5 incorrectly inferred residues, TT_AA_ML_JTT had 5 incorrectly inferred residues and TT_AA_ML_JTT+G had 5 incorrectly inferred residues. Amino acid sequences aligned using CLC bio v. 4.1.2.

It is interesting that the JTT and JTT+G models used with the MrBayes topology reconstructed identical sequences while the JTT and JTT+G models used with the true tree topology reconstructed sequences with the least amount of sequence identity among all inferences. This finding suggests that the gamma distribution makes a substantial difference considering that the gamma parameter is the only variable that differs. TT_AA_ML_JTT is the most accurate inference having only four incorrectly inferred residues: E121K, S152P, N178K, and K194I. TT_AA_ML_JTT+G is the least accurate inference containing seven incorrectly inferred residues: D94E, E121K, R125C, S152P, D174E, N178K, and K194I. So far, we have only witnessed cases where tree topology is not so informative, however, in this instance, the same models used with different tree topologies, namely the true tree and MrBayes tree, have a considerable effect on ASR inferences.

jModelTest selected the best model to be JTT+I which does not include gamma. However, PAML does not offer models that include invariable sites parameters, so our next available option was the JTT+G model. Many uncertainties could be resolved if we were able to construct ancestral sequences under the absolute best-fit models provided by jModelTest and ProtTest [8]. It would also be interesting to further dissect the role of topology in this instance. Since the MB_AA topology is less accurate that the MB_DNA topology when compared with the true tree, it would be appealing to use the MB_DNA topology

with amino acid sequence input to infer the 3s.1 ancestral sequence under both the JTT and JTT+G models.

Topologies are also not so informative under the DAYHOFF model in this study when considering the ancestral sequence inference. Both MB_AA_ML_DAYHOFF and TT_AA_ML_DAYHOFF sequences have 5 incorrectly inferred residues, however, these sequences are not identical (Figure 6.10)

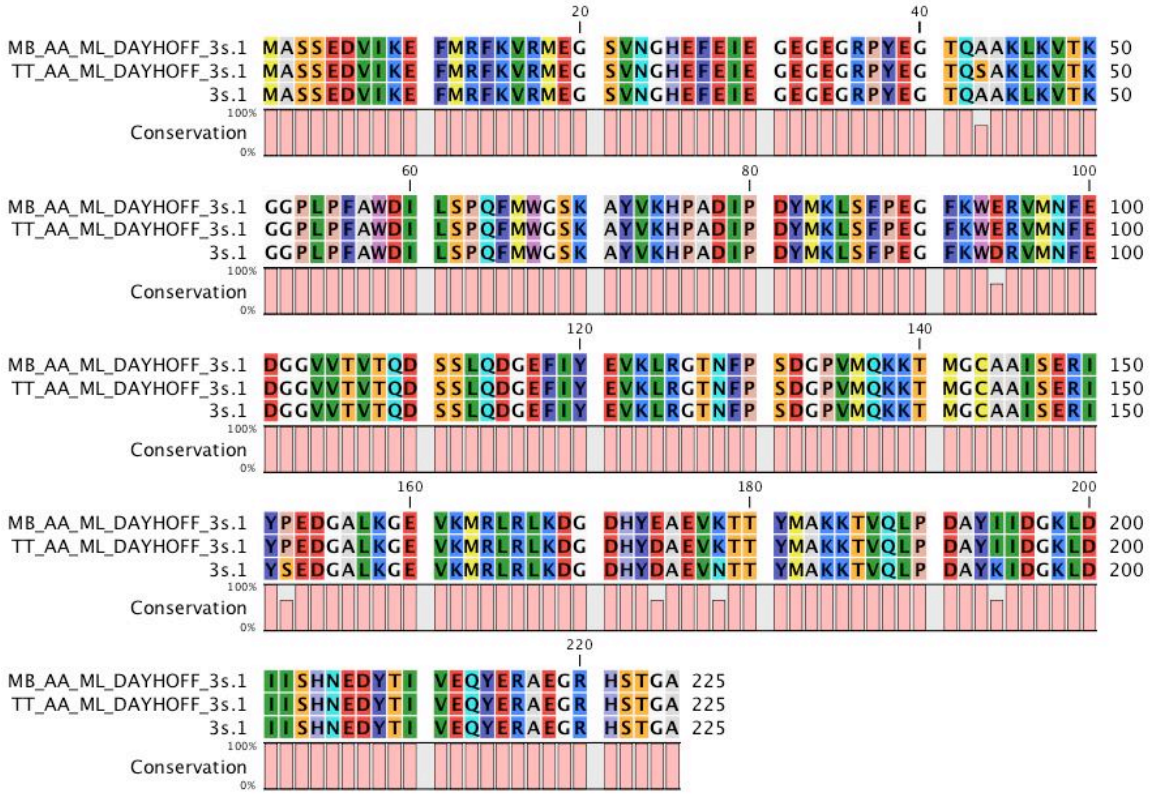



**Figure 6.10. Multiple sequence alignment of 3s.1 and Dayhoff acid inferences.** Multiple sequence alignment of MB_AA_ML_DAYHOFF and , TT_AA_ML_DAYHOFF. Both had five incorrectly inferred residues, and they share four of these residues (D94E, S152P, N178K, and K194I) and each differs by the fifth residue: MB_AA_ML_DAAYHOFF A43S and TT_AA_ML_DAYHOFF D174E. Amino acid sequences aligned using CLC bio v. 4.1.2.

Comparing Sequence Reconstructions

All ML and MP inferred sequences considered together contain a total of ten incorrectly inferred sites. All sequences incorrectly inferred residues S152P and K194I. Replacements at these residues have never been shown to result in a color change event and seem to only play a neutral role in FP evolution. Amino acid sites 152 and 194 do not directly lie within the chromophore environment when viewing their location on the mRFP structure (Figure 6.11).

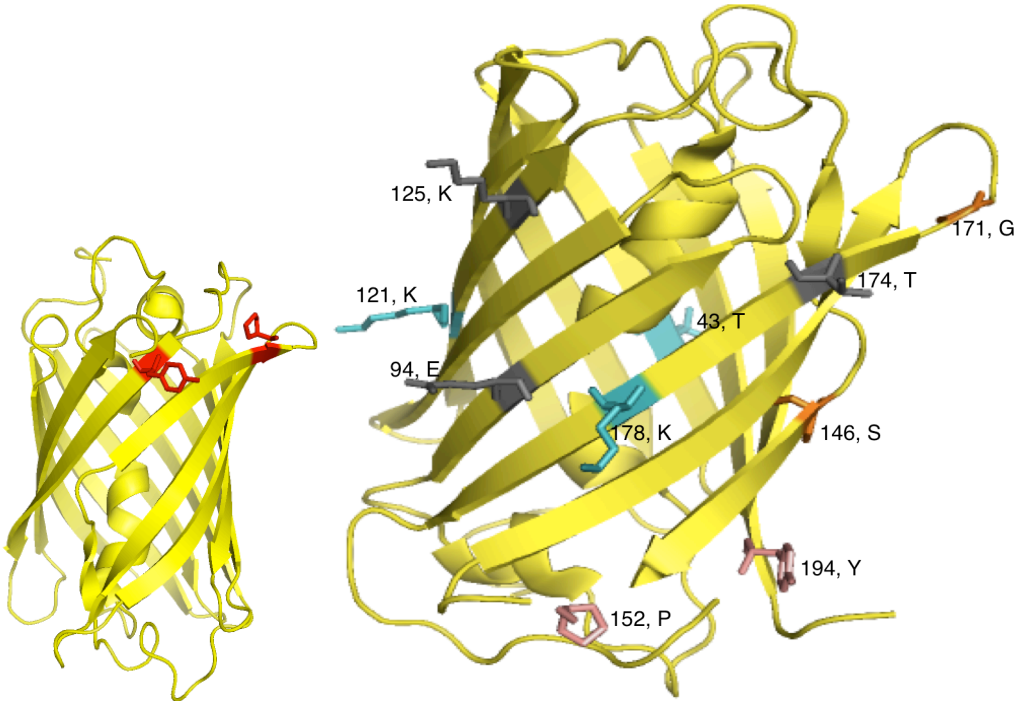



**Figure 6.11 Incorrectly inferred residues.** Structure on left is the mRFP 1.0 protein structure with only residues 152 and 194 highlighted in red. Structure on the right is rotated compared to the image on the left and highlights all 10 sites where incorrectly residues are inferred. Pink residues represent sites 152 and 194.

The type of evolutionary model selected for in ML analyses plays an important role in sequence inferences in this study. All models infer unique sequences that are almost never identical to a sequence inferred under a different model except in the case of AA_ML_JTT+/-G. Use of topology plays a less influential role over the evolutionary model. Cases when topology made a difference in ASR inferences were found among DNA parsimony inferences and amino acid inferences. DNA inferences retrieved identical sequences under the GTR+G and K80 models despite the topology used. Based off of these sequence reconstructions alone, the MB_AA_JTT sequence inference is the most accurate and the MB_AA_JTT+G is the least accurate. Assessing the accuracy of the different ASR analyses thus far is difficult since both the best and worst inference lie in the same category.

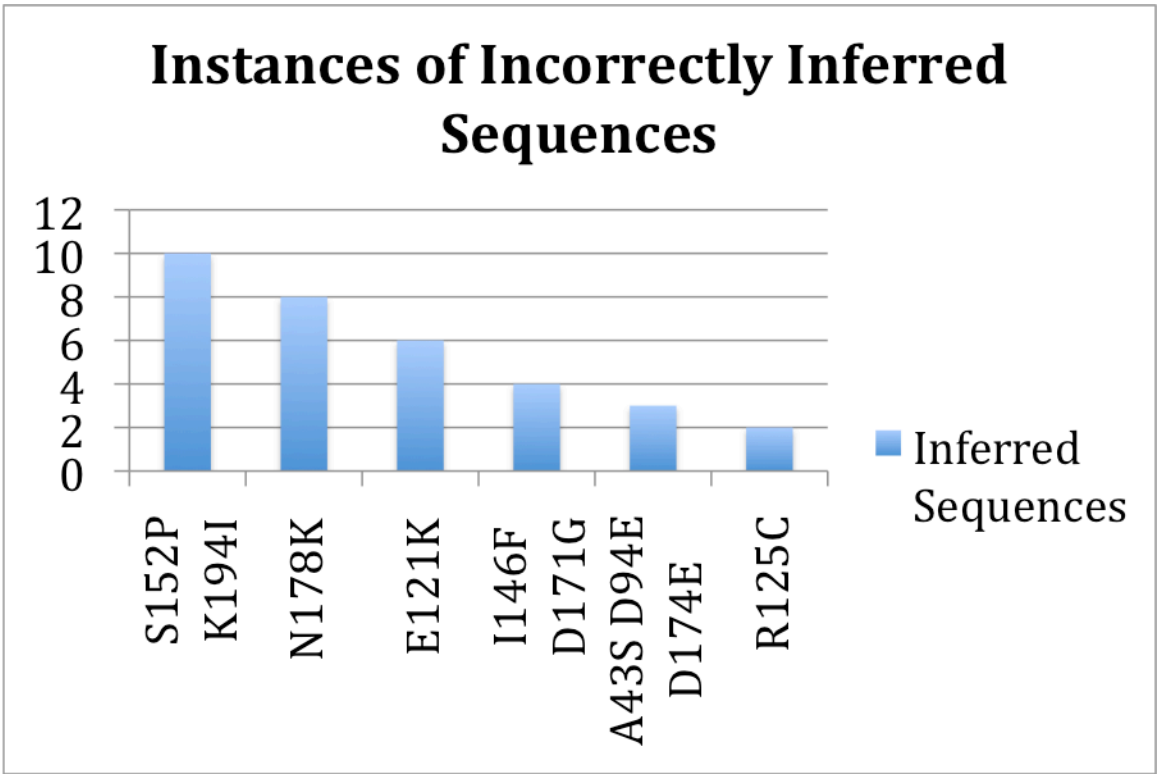



**Figure 6.12. Number of inferences with incorrectly inferred residues.** Bar graph shows the number of ancestral 3s.1 sequence inferences that incorrectly inferred a particular residue. Y-axis is the number of inferred sequences and the X-axis designates the amino acid site and replacement.

## Protein Resurrections

The next big question of course is whether any of these incorrectly inferred sites from the ASR analyses affect the phenotypes displayed by the ancestors when the incorrect residues are incorporated into their respective encoded proteins. We resurrected the incorrectly inferred ancestors in the laboratory to determine what, if any, effects these incorrectly inferred sites have on the true ancestral phenotypes associated with node 3s.1.

3s.1 is a dim yellow fluorescent protein variant with an emission peak at 528 nm and a minor peak at 570 nm (Figure 6.13.)

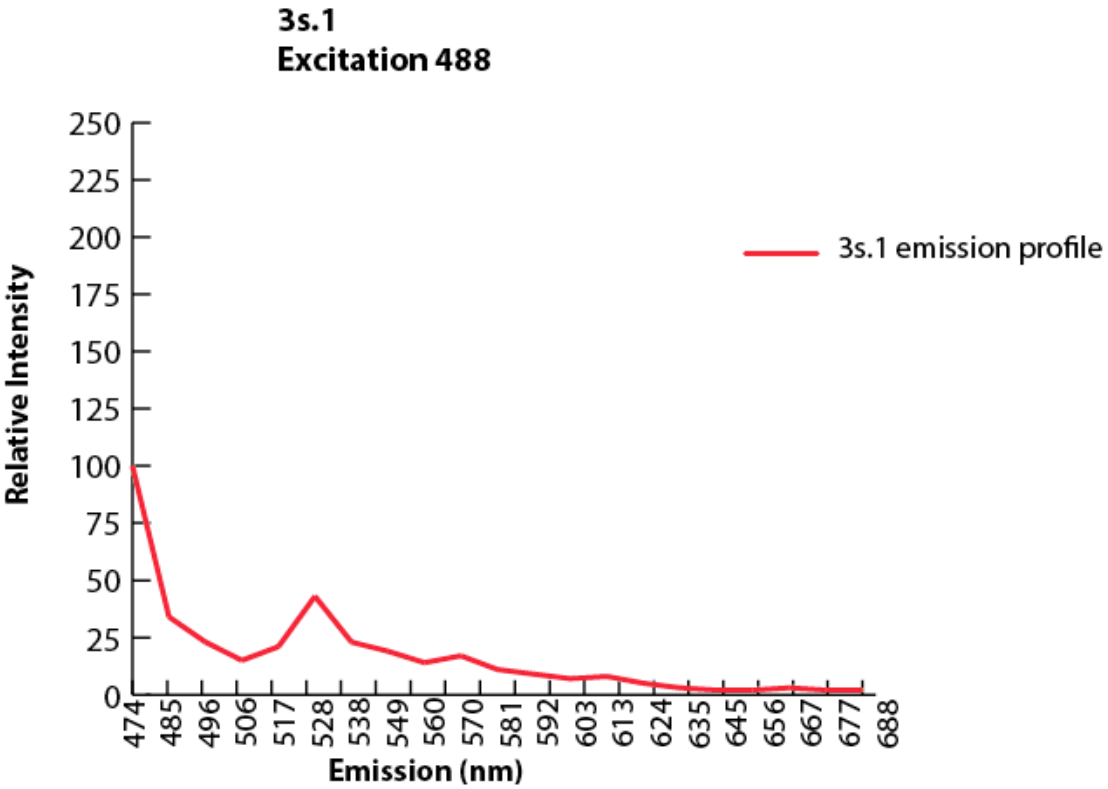



**Figure 6.13. 3s.1 spectra.** When 3s.1 is excited at 488 nm, it emits at a major peak 528 nm and a minor peak at 570.

Results and Discussion

All incorrectly inferred 3s.1 ancestral sequences display phenotypes very different from the true ancestor (Figure 6.14 and Table 6.1). All inferences are brighter and have emission profiles dissimilar to 3s.1 and seem to represent yellow and orange phenotypes.

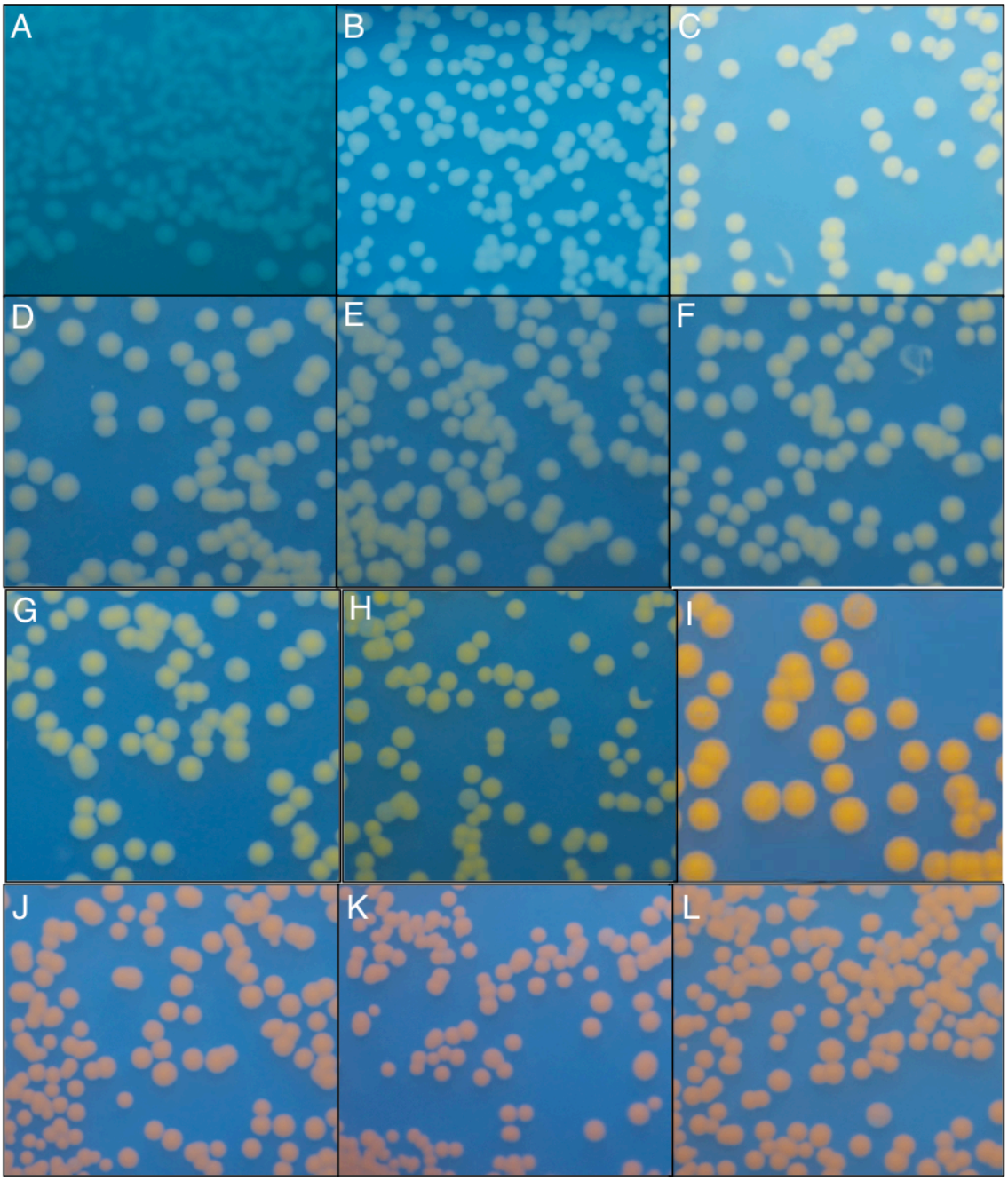



**Figure 6.14. Phenotypes of ancestral proteins**. Bacterial colonies expressing fluorescent proteins on an LB/CARB/IPTG/Agar plate under ultra violet light (365 nm). (a) pET-15b. (B) 3s.1. (C) TT_AA_ML_JTT. (D) TT_AA_MP. (E) TT_AA_ML_DAYHOFF. (F) MB_AA_ML_JTT+G. (G) MB_AA_ML_DAYHOFF. (H) TT_AA_ML_JTT+G. (I) TT_DNA_MP. (J) TT_DNA_ML_K80. (K) MB_DNA_MP. (L) TT_DNA_ML_GTR+G.

**Table 6.2. Phenotypes of 3s.1 and ASR 3s.1 inferences.** An image of the protein expressed within a bacterial colony are shown and the proteins maximum emission peak is listed.

| | Color | Name of Protein | Em MAX (nm) |
|---|---|---|---|
| A |  | (Control - No FP) | n/a |
| B |  | 3s.1 | 528 |
| C |  | TT_AA_ML_JTT | 570 |
| D |  | TT_AA_MP & MB_AA_MP | 570 |
| E |  | TT_AA_ML_DAYHOFF | 570 |
| F |  | MB_AA_ML_JTT+G<br>MB_DNA_ML_JTT | 570 |
| G |  | MB_AA_ML_DAYHOFF | 570 |
| H |  | TT_AA_ML_JTT+G | 570 |
| I |  | TT_DNA_MP | 570 |
| J |  | TT_DNA_ML_K80 | 570 |
| K |  | MB_DNA_MP | 570 |
| L |  | TT_DNA_ML_GTR+G | 570 |

Since the TT_AA_ML_JTT inference had the least amount of incorrectly inferred residues when compared with the true 3s.1 ancestor, we were curious to see if the phenotype was most similar to 3s.1 as well.

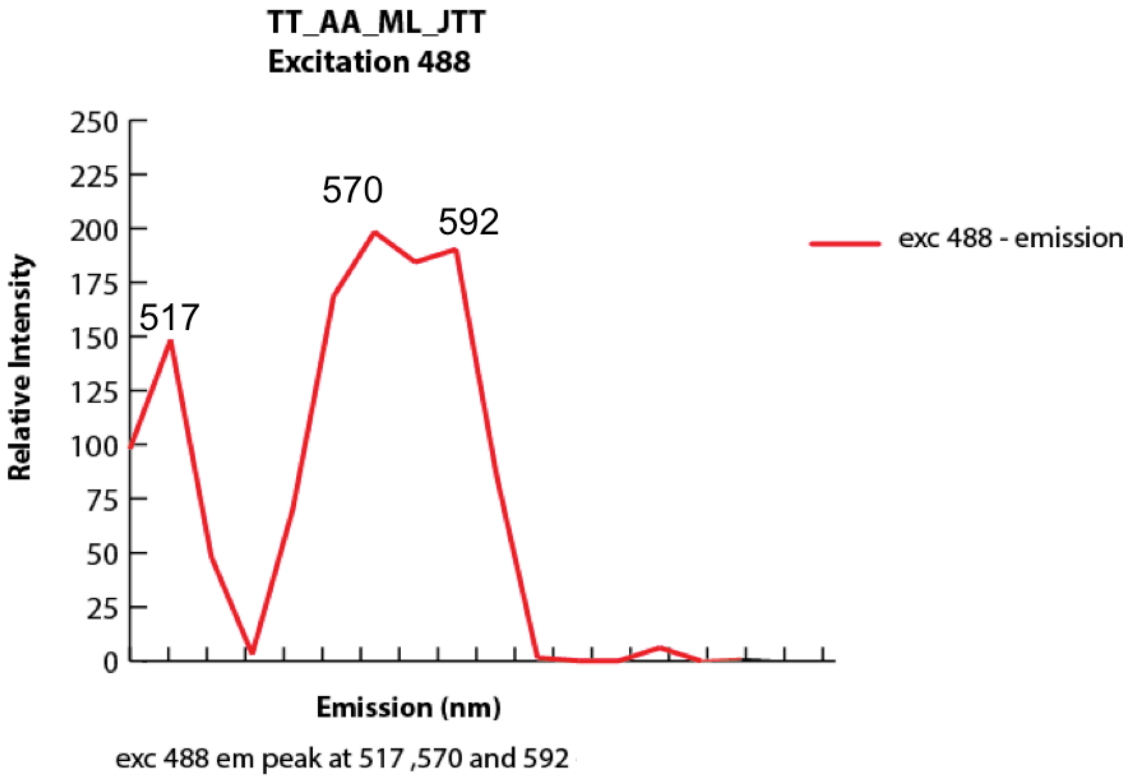



**Figure 6.15. TT_AA_ML_JTT spectra.** Spectra of TT_AA_JTT under confocal microscope excited at 488 represented by red line.

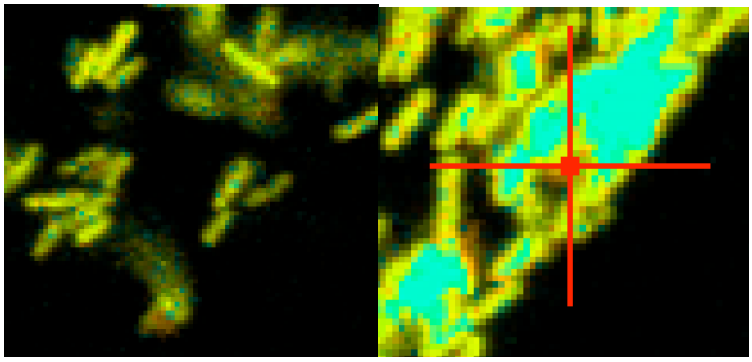


**Figure 6.16. TT_AA_ML_JTT fluoresence captured under microscope.** Bacterial colonies expressing TT_AA_ML_JTT visualized under UV 365 nm show a bright yellow phenotype (Figure 6.14 Table 6.1). Next, we were curious what the spectra for the least similar ancestral sequence inference TT_AA_ML_JTT+G with 7 mutations looked like.

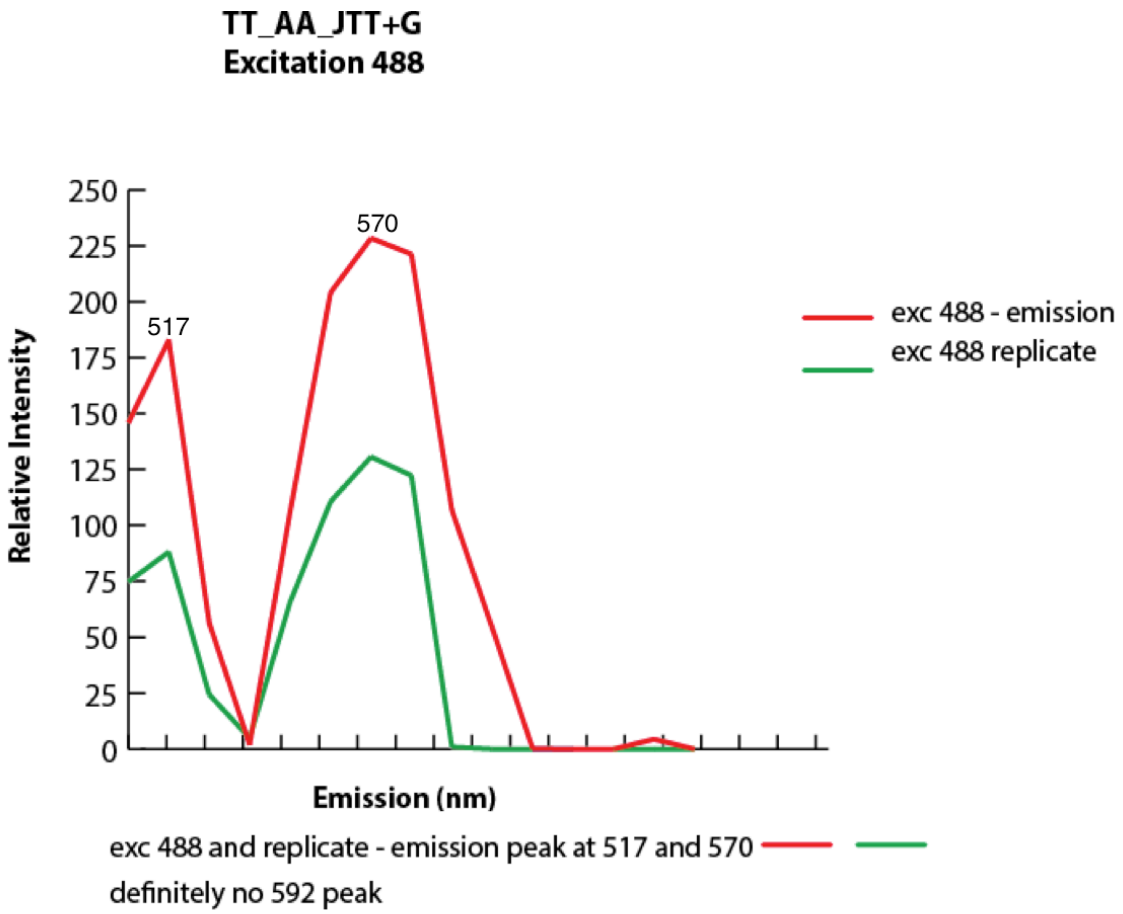



**Figure 6.17. TT_AA_ML_JTT+G spectra.** Spectra of TT_AA_JTT+G under confocal microscope excited at 488 represented by both red and green lines.

Despite these results being preliminary, they highlight the fact that ancestral sequence reconstruction in prone to error depending on the type of analyses. Both parsimony and likelihood led to incorrect ancestral phenotypes. Further, within the likelihood approach, different models of evolution influenced the inference of ancient sequences. In total, the experimental phylogeny is already proving useful to the field of ancestral sequence reconstruction by providing insights into the accuracy of different approaches which in turn will help guide the field in the coming year

## References


1. Bull, J.J., et al., *Experimental Molecular Evolution of Bacteriophage-T7.* Evolution, 1993. **47**(4): p. 993-1007.
2. Hillis, D.M., *Approaches For Assessing Phylogenetic Accuracy.* Systematic Biology, 1995. **44**(1): p. 3-16.
3. Hillis, D.M., et al., *Experimental Phylogenetics - Generation of a Known Phylogeny.* Science, 1992. **255**(5044): p. 589-592.
4. Hillis, D.M., et al., *Experimental Approaches to Phylogenetic Analysis.* Systematic Biology, 1993. **42**(1): p. 90-92.
5. Oakley, T.H. and C.W. Cunningham, *Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny.* Evolution, 2000. **54**(2): p. 397-405.
6. Posada, D., *jModelTest: phylogenetic model averaging.* Mol Biol Evol, 2008. **25**(7): p. 1253-6.
7. Yang, Z., *PAML 4: Phylogenetic Analysis by Maximum Likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.
8. Abascal, F., R. Zardoya, and D. Posada, *ProtTest: selection of best-fit models of protein evolution.* Bioinformatics, 2005. **21**(9): p. 2104-5.